The Role of Between-Case Effect Size in Conducting, Interpreting, and Summarizing Single-Case Research

Authors
William R. Shadish, University of California, Merced
Larry V. Hedges, Northwestern University
Robert H. Horner, University of Oregon
Samuel L. Odom, University of North Carolina, Chapel Hill

National Center for Education Research (NCER)
Meredith Larson (Project Officer)
Phill Gagné

National Center for Special Education Research (NCSER)
Kimberley Sprague

December 2015

NCER 2015-002
U.S. DEPARTMENT OF EDUCATION

INSTITUTE OF EDUCATION SCIENCES

This paper was prepared for the National Center for Education Research, Institute of Education Sciences under Contract ED-IES-12-D-0015. Meredith Larson was the project officer.

Disclaimer

The Institute of Education Sciences at the U.S. Department of Education contracted with Westat to develop a paper on the use of effect sizes in single-case design research. The views expressed in this report are those of the authors, and they do not necessarily represent the opinions and positions of the Institute of Education Sciences or the U.S. Department of Education.

December **2015**

Shadish, W.R., Hedges, L.V., Horner, R.H., and Odom, S.L. (2015). The Role of Between-Case Effect Size in Conducting, Interpreting, and Summarizing Single-Case Research (NCER 2015-002) Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. This report is available on the Institute website at http://ies.ed.gov/.

Alternate Formats

Upon request, this report is available in alternate formats such as Braille, large print, audiotape, or computer diskette. For more information, please contact the Department's Alternate Format Center at 202-260-0852 or 202-260-0818.

Disclosure of Potential Conflict of Interest

Westat Inc. is the prime contractor for the NCER Analysis and Research Management Support project, under which this paper was developed. Tamara Nimkoff was the project manager. The authors of this paper are not aware of any conflicts of interest.

Acknowledgments

The authors thank the members of the Technical Advisory Group (Ann Kaiser, Vanderbilt University; Thomas Kratochwill, University of Wisconsin Madison; Kathleen Lane, University of Kansas; Daniel Maggin, University of Illinois, Chicago) for their helpful comments on a previous draft of this paper. Of course, remaining problems and errors are the responsibility of the authors.

Contents

List of Tables

List of Figures

1. Overview and Purposes

Educators are increasingly committed to adopting practices that are "evidence based." Most education researchers agree that the randomized experiment[1] has important advantages as a method for identifying practices that are evidence-based yet also recognize that randomized experiments are not always feasible. In such cases, those researchers will often choose another kind of experiment that is suitable to the task. For example, if the intervention is a remedial reading program given only to students who fail a reading test by scoring below a cutoff, a regression discontinuity design may be possible. Many such designs exist, some yielding weaker causal inferences, and others yielding stronger inferences (Shadish, Cook, and Campbell 2002). One of the stronger experimental design options is provided by single-case designs (SCD).[2]

Imagine an education researcher who wants to test whether a new treatment improves the social behavior of children with autism spectrum disorders (ASD).[3] One option is to randomly assign children to receive either the new treatment or some comparison condition like a wait-list control or a standard treatment. Several problems may arise in this situation. A well-powered randomized experiment will usually require more children with ASD than are available to most researchers because of the relatively low prevalence of the disorder. Also, if the treatment is classroom or school-based, the random assignment may occur at the class or school level, which then necessitates a study that sometimes exceeds the resources available to most researchers who do not have substantial grant funding (Odom and Lane 2014). In addition, both parents and researchers may be concerned about placing children in a research design that requires withholding educational supports (e.g., in a control group) for an extended period of time (especially for young children in critical developmental periods).

---

[1] This term and others in similar font throughout the document are hyperlinked to a glossary at the end of the document. Click on the term, and it will take you to the glossary definition.

[2] Much of this paper will apply equally well to research outside education such as psychology, social science, and medicine, where short time series like SCDs are widely used. In medicine, for example, SCDs are often called N-of-1 trials.

[3] In this example, the case is the child. Cases are usually individuals like children or teachers but can also be aggregates such as classrooms.

In such cases, the researcher may appropriately choose to use an SCD in which the child is observed frequently over time, both in the absence or presence of an experimenter-controlled treatment manipulation. Roughly speaking, the logic of SCDs is that treatment effectiveness is inferred if the outcome covaries as expected in the presence and absence of the manipulated treatment. This characterization oversimplifies professional standards for valid causal inference in SCD research, so we return to this point in section 2.2.3. When certain conditions are met, such designs can yield a strong causal inference about whether the treatment works, and they can do so with a far smaller number of cases than would be needed to conduct a between-groups experiment (e.g., a randomized experiment).

In principle, then, SCD studies could contribute to evidence-based practice reviews about what works when they address the question asked in the review, but they are omitted in such reviews more often than would seem warranted given the quality of evidence they yield. A key reason is that SCD researchers historically have emphasized visual analysis to assess and report the effects of treatments and they tend not to report statistical analyses. Skepticism about statistics among SCD researchers is founded partly in doubt that many statistical analyses that could be applied (e.g., regression, overlap statistics, standardized effect sizes) successfully capture the nuances of SCD data (e.g., most overlap statistics do not take trend or autocorrelation into account; most trend is assumed to be linear), and partly in the fact that many of those statistics were or still are under development, with no consensus on why one should be chosen over another. Among the many possible statistical analyses that could be used, standardized effect sizes[4] are particularly valued because they let reviewers compare results across studies having different outcome measures that otherwise could not be easily compared. As per our definition in the glossary, we use this term in its most general sense—any effect size that puts measures on a common metric, that is, a scale that purports to have the same meaning over studies. Well-known examples are the standardized mean

[4] When we mention standardized effect sizes, many readers may assume we are talking about evidence-based practice reviews that include a meta-analysis. Of course, standardized effect sizes are essential to meta-analysis in most cases. Yet even when an evidence-based practice review does not include a meta-analysis, standardized effect sizes are still useful to allow reviewers to see whether effects on different outcomes within the same study have the same magnitude. In addition, a certain class of those effect sizes that we call "between-case effect sizes" also allows the reviewer to see whether effects from SCDs are of the same magnitude as effects from studies using between-groups designs. The example we give in section 6.1 of this paper about how effect sizes are used in a recent What Works Clearinghouse (WWC) review of the Spelling Mastery curriculum (U.S. Department of Education 2014b) will clarify this point. That review was not a meta-analysis, but it did report standardized effect sizes for the purposes just stated.

difference statistic and the correlation coefficient, both of which are a special case of our more general usage that also includes odds ratios, risk ratios, and the overlap statistics in SCDs.

Not all effect sizes are standardized. An example is the simple raw difference between two means, often appropriate when all studies use exactly the same outcome (e.g., dollars, days in the hospital) so that there is no need to standardize this raw mean difference (Shadish and Haddock 2009). In SCD research, the difference between two phase means is, in fact, an effect size, albeit not standardized. However, such raw mean differences are not standardized (not on a scale with a mean of zero and standard deviation of one). Further, some effect sizes put diverse outcomes on a common metric but without standardizing. An example is an odds ratio in which dichotomous outcomes of different sorts can be put into a common effect size metric, but that metric is not standardized in the sense this term is used in statistics—dividing a value by its standard deviation to yield a score that has a mean of zero and a standard deviation of one.

Adding standardized effect sizes may seem to some to be a big step. It is worth remembering, though, that it was not so long ago that between-groups studies failed to report standardized effect sizes, too, but they are now commonly reported. Accepted methods for calculating standardized effect sizes of SCD studies already exist and could have the beneficial effect of making it easier to interpret SCD studies in syntheses, reviews, and meta-analyses focused on identifying evidence-based interventions. SCD research already makes many high-quality contributions to our knowledge of what works. The single best thing we can add to existing practice to make sure this research is accessible and influential is to have good effect size measures. We propose that a class of such effect sizes, which we will call standardized between-case effect sizes (see section 3.1.2), are the most applicable for this purpose. Hence, the goals of this paper are to suggest appropriate measures of between-case effect size that can be applied to SCD research and to suggest how SCD researchers can report data from individual studies to make it easier for others to compute effect sizes in the future.

Visual analysis has a long history and conceptually compelling set of rationales buttressing its use, and nothing in this paper is intended to challenge the continued thoughtful use of visual analysis in SCD research. Nonetheless, failure to report statistical analyses in SCDs is an obstacle because many researchers conducting evidence-based practice reviews value conventional forms and

representations of evidence using statistics to integrate findings across multiple studies (U.S. Department of Education 2014a).

1.1 Who Are the Audiences for This Paper?

This paper has four primary audiences and different parts of this paper will be of primary interest to different audiences. The first is SCD researchers who conduct the original SCD studies. We hope to add a tool to the repertoire of single case researchers that may be useful in its own right and that may help highlight the value of SCD studies to those outside the SCD community. The second is those who review literatures to identify what works. Sometimes those reviewers are SCD researchers themselves, but just as often the reviewers are scholars outside the SCD community in academia, private contract research firms, or governmental bodies at all levels that have interests in questions about effective interventions. For all those reviewers, we aim to provide statistical recommendations about effect size calculations and their use in meta-analysis and pertinent examples to encourage the inclusion of SCD studies in evidence-based practice reviews. The third audience is those in policy positions with resources and interests in improving the use of SCDs both as a method in their own right and as one of many methods that can contribute to knowledge of what works. For them, this paper, especially Chapter 6, includes numerous suggestions for future directions that can help advance these ideas. The fourth audience is scholars with a statistical and methodological bent who might themselves help make those advancements. Right now, relatively few of them are involved in such efforts, but the intellectual issues in the analysis and meta-analysis of SCD research are theoretically interesting in their own right (Shadish 2014a).

1.2 An Overview of the Paper

Readers may find it useful to see a brief summary of both the structure of the paper and of its recommendations. Its structure is as follows:

- Chapter 2 defines SCDs in detail, gives examples of their basic forms, and discusses the logics that buttress causal inference in SCDs.
- Chapter 3 introduces the nature of standardized effects sizes, discusses how and why they can help get SCD research included in evidence-based practice reviews, shows why the distinction between what we will call *within-case* and *between-case* effect sizes is important for

evidence-based practice, and reviews the strengths and weakness of three available between-case effect sizes for SCDs.

- Chapter 4 suggests an approach to reporting between-case effect sizes in SCD research and a discussion of relevant assumptions. In addition, we recommend that raw data from SCD studies be made more readily available in tables or as supplementary materials for extended analysis (including meta-analyses) by others, now and in the future.
- Chapter 5 describes some other benefits that will ensue as SCD researchers begin to use between-case effect sizes more widely, including (a) the ability to design SCDs that are more sensitive to detect an anticipated effect size, (b) new kinds of studies examining assumptions about SCDs such as differences between visual and statistical analysis, (c) exploration of any relationship that might exist between effect size and researcher decisions to publish SCD results, and (d) accumulation of descriptive data about the size of the effects across many content areas using SCDs as well as within each content area to see what constitutes a small, medium, or larger effect.
- Chapter 6 (a) shows how to use between-case effect sizes to identify evidence-based practices in reviews that include SCDs and includes an example borrowed from the What Works Clearinghouse (WWC),(b) presents an introduction to how such effect sizes can be used with modern meta-analytic methods to provide more reliable and valid summaries of evidence, and (c) highlights conceptual issues that will necessarily arise when a reviewer wants to consider including results from both SCDs and between-groups studies in a single review.
- Chapter 7 discusses the kinds of short-term and long-term activities that might eventually foster a consensual collaboration among SCD researchers and statisticians to encourage not only effect size reporting but possibly also many other methods of statistical analysis.

These chapters make the following suggestions for the reporting of standardized between-case effect sizes in single-case designs:

1. The chance that reports of results from SCDs will be included in evidence-based practice reviews will increase if those reports include a standardized between-case effect size because
    a. Standardization allows comparing results from outcome measures in different metrics; and

   b. Between-case effect sizes are in the same metric as the effect sizes used in between-groups studies, allowing comparison of results from SCD to those from between-groups studies.[5]
2. Three approaches for calculating standardized between-case effect sizes for SCDs currently exist, each with different strengths and weaknesses. The potential user will need to read about these strengths and weaknesses before deciding whether to use any or all of them.
3. Reports of SCD results will be strengthened when the following features are included:
   a. A standardized between-case effect size, the standard error of the effect size, and an associated inferential test, for each reported outcome;
   b. Citation to the exact effect size used and associated software or syntax; and
   c. A list of assumptions on which the use of the effect size is predicated, with statistical tests or other discussion of the plausibility of those assumptions.
4. In addition to presenting their data in graphic form, SCD researchers should have their raw data available in tables (e.g., in a spreadsheet) that may be included in an appendix or repository so other researchers can access the original data. This will ameliorate problems caused by the time consuming and sometimes erroneous digitation of data from the publically available graphic forms.
5. Meta-analyses that include SCD research would benefit from use of a more extensive set of modern meta-analytic statistics than has generally been the case in the past.

This paper elaborates these key points. But first, we introduce the reader to the general nature of single-case experimental designs as they are used to assess the effects of interventions.

[5] Notice this claim is not that results from SCDs and between-groups studies should be combined, just because the effect size allows them to be combined. Section 6.3 of this paper discusses many issues related to whether such combining should be done.

2. Background: Single-Case Designs

SCDs are widely used to test the effects of interventions in many parts of education, psychology, and medicine (Gabler, Duan, Vohra, and Kravitz 2011; Horner and Odom 2014; Shadish and Sullivan 2011; Smith 2012). They are used in such education-related fields as school psychology, special education, remedial education, early intervention, occupational therapy, and speech pathology (Shadish and Sullivan 2011; Smith 2012).

The core features of SCDs when used to test the effects of interventions include the repeated assessment of an outcome over time (i.e., a time series) within a case (which could be a child, a classroom, etc.), both in the presence and absence of an intervention, where the experimenter controls the timing of the intervention both within and across cases (Horner et al. 2005; Kratochwill and Levin 2014) (i.e., an interrupted time series). For example, most publications that use SCD methodology report results for three or more cases, some of which may receive the intervention at different times or may have the intervention removed and reintroduced over time (Shadish and Sullivan 2011). In all their forms, SCDs are indeed experiments (albeit not necessarily randomized experiments) because they meet the classic definition of an experiment throughout the history of science, i.e., that the researcher deliberately introduces a treatment to discover its effects, using any number of additional methods to help control for confounding variables (Shadish, Cook, and Campbell 2002).

Sometimes, the SCD researcher randomly assigns conditions to time (Kratochwill and Levin 2010). For example, Himle, Woods, and Bunaciu (2008) treated four children with tics. They presented one of three conditions in each session (two treatments or no treatment), with the particular condition assigned to session randomly. Other examples of SCDs that used random assignment of conditions to time are Conelea and Woods (2008), Codding, Livanis, Pace and Vaca (2008), and Glover, Roane, Kadey, and Grow (2008). Kratochwill and Levin (2010) describe many variations of how random assignment of conditions to time might occur. When such assignment occurs, the result is a randomized experiment with the usual benefits of randomization for internal validity (Reichardt 2006), albeit where the counterfactual is within-case rather than between-case (Shadish, Hedges, Pustejovsky, Rindskopf et al. 2014).

Indeed, one of Sir Ronald Fisher's first examples of a randomized experiment—the famous Lady Tasting Tea experiment (Salsburg 2001)—was exactly this kind of randomized SCD. A lady in a faculty club said she could taste the difference between cups in which the milk was poured first and tea second, versus tea first and milk second. Fisher arranged to present eight cups of tea to her, half randomly using one order of pouring and half with the other. Salsburg (2001) reports the lady chose correctly every time. Fisher later showed how to analyze this as a randomized experiment. The example is particularly pertinent since Fisher is rightly understood to be the foremost developer of between-groups randomized experiments, so it is worth recalling his primary interest was in randomization, whether in a within-case or between-groups context. In medicine, where SCDs are called N-of-1 studies, such random assignment of condition to time is common, although those medical N-of-1 studies differ from social science SCDs in other methodological features that would warrant additional study in a paper that thoroughly compared them. For present purposes, however, because such random assignment is rare in social science SCDs, this paper focuses primarily on the usual forms of SCDs that do not use random assignment but in which replication is used to establish plausible causal inference through repeated demonstrations of an experimenter-controlled cause and effect relationship.

We assume in the following discussion that the reader has some familiarity with SCDs; readers who wish more background can refer to standard texts in the field (e.g., Barlow, Nock and Hersen 2009; Gast and Ledford 2014; Kazdin 2011; Kratochwill and Levin 2014). The presentation of these designs focuses on the general methodological categories of SCDs, not particular implementations of them that may or may not meet acceptable professional standards. Consistent with this intent, we provide hypothetical examples of each design to clarify its basic features. In addition, we greatly oversimplify description of the treatments for didactic purposes, recognizing that the treatments used in SCD studies are often complex, sophisticated, and theory driven. Finally, we focus only on the use of SCDs to identify effective treatments or components of treatment, though SCDs can be used for other purposes.

2.1 Basic Forms of Single-Case Designs

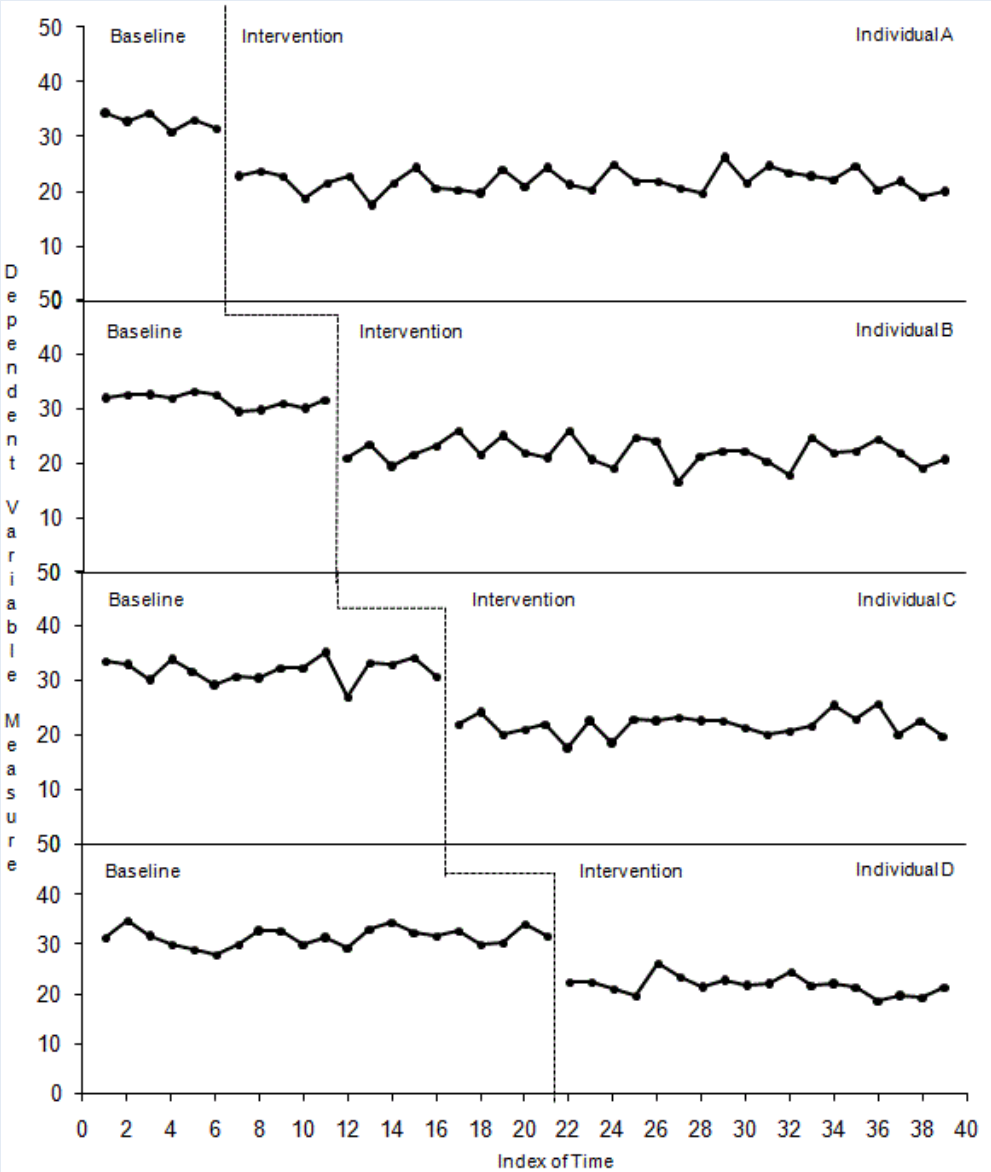
SCDs typically come in one of four basic forms or some combination of those four forms (Gast and Ledford 2014; Kazdin 2011; Shadish and Sullivan 2011; Smith 2012). This section is intended to be pedagogical, and so uses simple, hypothetical examples that often present clear effects so that the logic and concepts that buttress SCD research are easily seen. Examples in subsequent sections of this manuscript use real data with more real world complexities.

2.1.1 Multiple Baseline Design

By far the most common SCD is the multiple baseline design (Shadish and Sullivan 2011). The term *multiple baseline* refers to the experimenter monitoring multiple cases of an intervention with all cases being observed on the outcome at the same start time, but deliberately delaying the introduction of treatment at different times for different cases, and then following all cases through to a common end point. Each case has (at least) two phases, baseline and treatment, where the word *phase* refers to a time period during which observations are made while a constant intervention occurs, and those phases change over time. The length of the delay in a multiple baseline design (i.e., how long to wait after the first case gets treatment before the next one does) is usually determined by a participant's (case's) performance. The delay should be long enough for the treatment to show a stable behavior pattern in the treatment phase for the first case before moving on to the second case; treatment onset for the third case should occur after the performance of the second case demonstrates a stable pattern, and so on, if there are more than three participants.

Example 2.1. A Multiple Baseline Design. In this hypothetical study, the researcher wants to give the same intervention to each of four individuals, with the goal of reducing an undesirable outcome. For example, this could involve a treatment to reduce off-task behavior by young children with intellectual disabilities. To meet the standard of allowing multiple opportunities for the treatment to show an effect, the researcher gives the intervention at different times for each individual—at time 7 for Individual A, time 12 for Individual B, time 17 for Individual C, and time 22 for Individual D.

Multiple Baseline Design



The graph presents the results of the study, and its structure is typical of how SCD researchers present study results. The outcome is on the vertical axis, and time is on the horizontal axis. Each dot in the body of the graph is an outcome observation (that itself might be a composite of multiple observations within that one time point). The vertical dotted lines tell us when the intervention was given for each case. The results in this example suggest that the frequency of off-task behavior did decline after the introduction of treatment, but not before, and that decline occurred at four unique times that coincide with when the treatment was introduced to the four cases. It is difficult to think

of a plausible alternative explanation for why that decline would occur at four different times for four different individuals at just the time when treatment occurred.

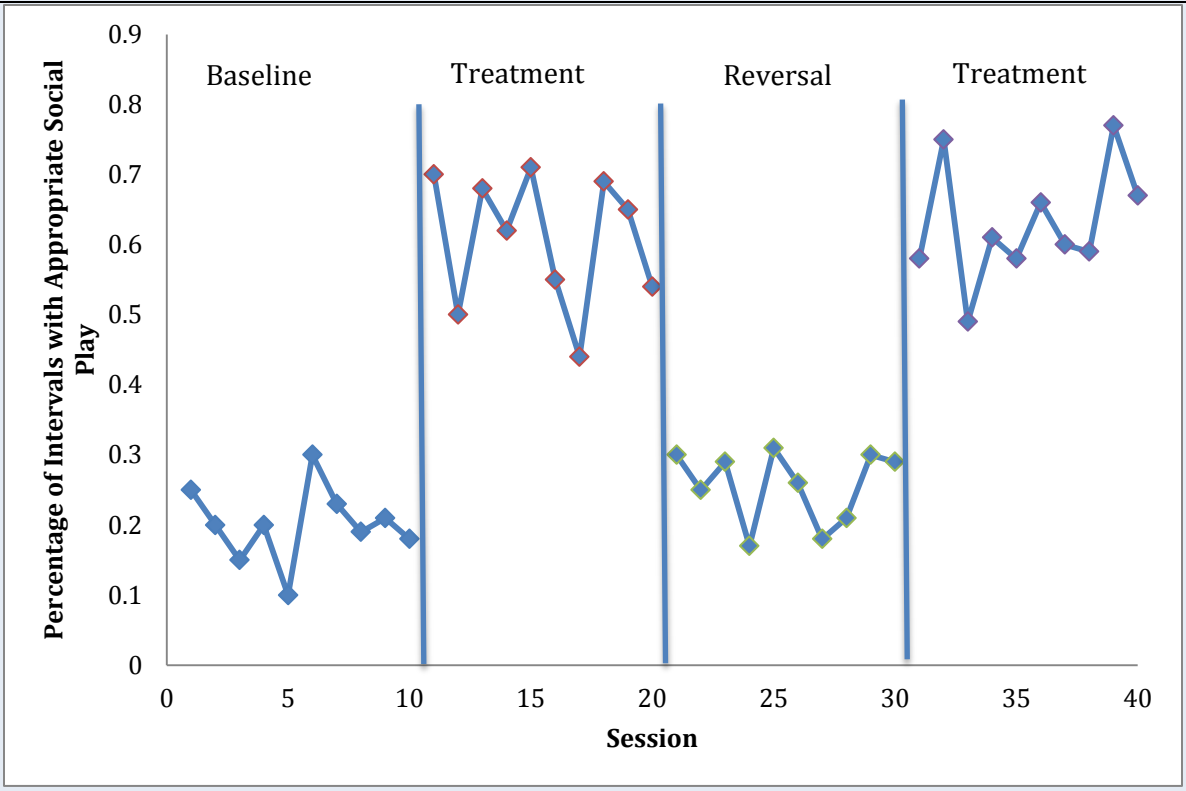
The multiple baseline design highlights our earlier point that many variations exist on each of these basic types of SCDs. For example, the multiple probe design (Horner and Baer 1978) replaces continuous observation of outcome during baseline for each case with occasional observation of outcome at theoretically specified points. Further, Example 1.1 is technically an example of a multiple baseline design across participants, but the design can be done within a single participant across different settings, materials, or behaviors. Details of such variations are beyond the scope of this paper but are described by Gast and Ledford (2014).

2.1.2 Reversal Design

A second design is called a reversal design, or sometimes a withdrawal or an ABAB design (often seen as A-B-A-B). For example, an ABAB design might start with baseline observations in the absence of treatment (phase A_1), then introduce treatment (B_1), and then remove treatment (A_2) which is essentially returning to baseline, and finally reintroduce the treatment one more time (B_2). The second baseline phase (A_2) gives rise to the descriptor *reversal*, whereby the treatment is withdrawn and reverts to baseline.

Example 2.2. A Reversal Design. This example shows a treatment applied to just one person who is displaying inappropriate social behaviors (e.g., perhaps a child in a recess yard), and the example uses a phase reversal design. The outcome in this example is the percentage of intervals in which the person is observed displaying appropriate social behavior during each session. To oversimplify for pedagogic purposes, the design is established to determine if introduction of the treatment will be associated with an increase in appropriate social behavior. Good professional practice is to have enough time points (sessions) in each phase to allow documentation of a clear pattern (level, trend, variability). Documenting a within-phase pattern typically requires at least five time points in each phase, but the number of data points needed to document a stable pattern varies depending on the behavior of concern and the context. This example has 10 time points per phase, and this is helpful because more time points provide more information statistically and sometimes even clinically.

Reversal Design



Appropriate social behavior is low during baseline. When the treatment is introduced, appropriate social behavior increases, which is the first demonstration of a treatment effect. In the reversal phase, the treatment is withdrawn and appropriate social behavior decreases (a bad outcome); this is the second demonstration of a treatment effect. Finally in the last phase, the treatment is reintroduced and the outcome again improves, which is the third demonstration in which manipulation of the independent variable was associated with change in the dependent variable (appropriate social behavior) with effects that are consistent with the hypothesis that the treatment is functionally related to an increase in the level of appropriate social behavior.

Causal inference is supported by the demonstration that the outcome tracks the presence or absence of treatment when it is deliberately introduced or withdrawn multiple times. Of course, a reversal to baseline is only useful when the treatment does not have permanent effects. When that is not the case, for example, when the effects of a learning procedure cannot be "undone," a multiple baseline design will often be a more appropriate design choice.

Note that if this design only had two phases (AB) or even three (ABA), it would not meet the requirement of three opportunities to demonstrate a treatment effect outlined in some major statements of methodological standards in SCDs (e.g., Hitchcock et al. 2014; Horner et al. 2005; Kratochwill et al. 2010, 2013).

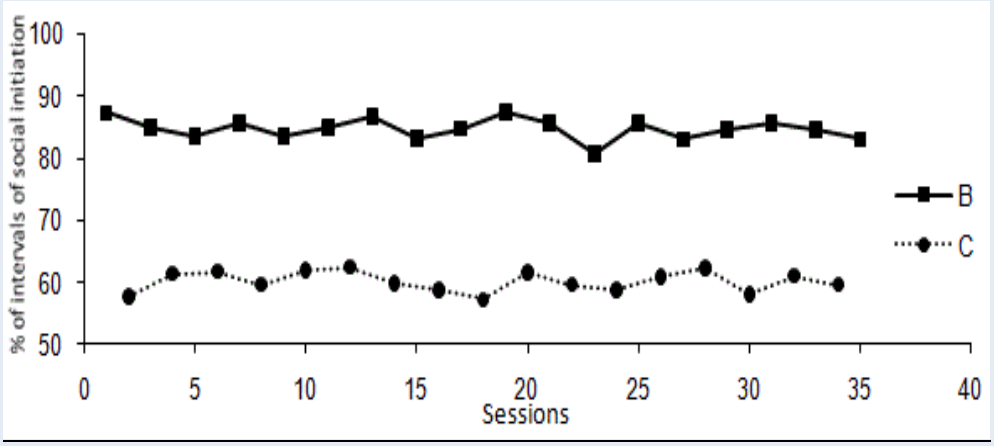
2.1.3 Alternating Treatments Designs

A third SCD is closely related to the reversal designs and is called the alternating treatments design (ATD). As its name implies, different conditions are rapidly alternated over time. For example, the order of treatments might be randomized across sessions, or treatment might be given in even-numbered sessions, with odd-numbered sessions being no treatment baselines. Phases in alternating treatment designs are very short, often just one measurement of the outcome, but those short phases are repeated many more times than in the usual reversal designs.

Example 2.3. An Alternating Treatments Design. This design might be used for a treatment to improve the percentage of time that a child with ASD initiated social behaviors. If the treatment were sufficiently fast acting to take effect quickly and sufficiently dependent on contingencies like a reward for social behavior, then treatment could occur on alternating sessions with the baseline being the other sessions. The order of alternating sessions is often randomly determined to control for order effects (e.g., the fact that baseline always occurs before the treatment condition might affect the results), and when not randomized, it is usually systematically counterbalanced. One can imagine many variations on this basic design, as when a single session might contain both B and C conditions, or B and C are assigned randomly to times.

The graph below shows the outcome data on the child's social behavior for two treatments, B and C, when B is applied in the first session, C in the second, etc. However, C could just as well be a baseline (no treatment) condition, or in addition to two treatments B and C, a baseline condition (A) could occur first, followed by B and C alternating treatments. In the graph below, the points representing outcomes on the child's social behavior of each treatment (B and C) are a separate time series of outcomes for each treatment. The outcomes for B are the black squares, and the outcomes for C are the black circles. To aid visual interpretation, the outcomes are connected by lines separately for each treatment condition.

Alternating Treatments Design



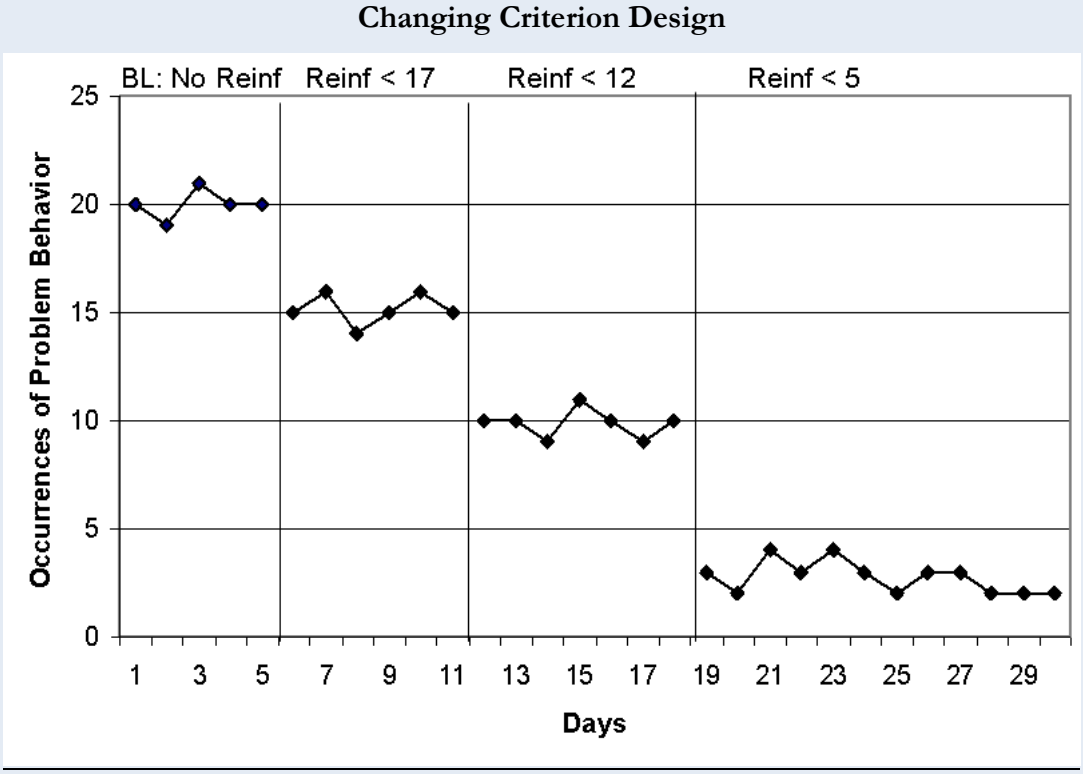
The researcher for this study would look at the difference in outcomes between B and C to determine if one treatment was having a consistently different effect from the other treatment. Clearly the outcomes under B are consistently higher than under C, suggesting that the B treatment is more effective at promoting social initiations. Support for such a causal inference is justified in a manner similar to the reversal designs, except the ATD presents many more demonstrations of the differential effects of treatments. The likelihood that any variable other than manipulation of the B and C conditions was responsible for change in social initiation is considered extremely low due to the predicted and consistent change in the outcome (social initiation) across the multiple B and C condition manipulations. Some ATDs also include a baseline condition as one of the alternative treatments. However, the research question addressed most often by ATDs is about differential treatment effects.

2.1.4 Changing Criterion Designs

The fourth and least frequently used of SCDs is the changing criterion design. The word *criterion* refers to a specified level of performance the case must attain to receive a reinforcement like a reward. In many treatments, such a reward is a key part of changing behavior, following well-known theories of operant learning. One might, for example, reward a child who is often disruptive in a classroom for remaining in seat and raising a hand to be called on. In many studies, the goal is to change the behavior quite substantially, for instance, virtually eliminating disruptive behavior.

However, attaining that great a change in a short period of time may not be realistic. Instead, the research initially targets simply lowering the existing level of disruptive behavior somewhat. Once the student reaches that level for multiple observations, then the criterion is changed to be more demanding. This process of changing the criterion continues until the student reaches the ultimate goal, for example, less than 5 percent observations with disruptive behavior.

Example 2.4. A Changing Criterion Design. In this example, a treatment is given to a child who is displaying problem behavior. During baseline, the child displays about 20 such behaviors each day. After baseline, the child is rewarded in the next phase for reducing problem behaviors to fewer than 17 per day. In the third phase, the criterion is reduced further to less than 12, and in the last phase to less than 5. At this point, the study may either end or continue with a generalization phase (we do not talk about such phases in this paper, but they are often a key part of demonstrating that treatment effects are lasting; Gast and Ledford 2014). The resulting data might be presented in a graph like this one, where the vertical lines indicate the time at which the criterion was changed.



Causal inference is supported by the demonstration that the behavior changes predictably in response to the experimenter-imposed manipulation of more demanding criterion for success.

2.2 Causal Inference in Single-Case Designs

The preceding discussion refers frequently to causal inference, that is, whether the treatment had an effect on the outcome. In modern experimental design and statistics, the most common theory of causation relies on the potential outcomes statistical model advanced by Donald Rubin (e.g., Imbens and Rubin in press; Rubin 1974; Shadish 2010), elaborated by many others (e.g., Morgan and Winship 2007), but first articulated by Jerzy Neyman in his master's thesis almost a hundred years ago (Neyman 1923). The definition of a treatment effect is then the difference between the outcome if a person receives a treatment and the outcome if that same person does not receive it—notice that, in theory, this effect is defined for each person. In practice, a person cannot both receive and not receive a treatment at the same time. Hence, much of experimental design is about how to estimate what would have happened to those who got treatment if they had not gotten treatment (often called the counterfactual inference). In a randomized experiment, for example, the source of counterfactual inference is the randomized control group, not a perfect estimate of the true counterfactual because the people in the control group are not the same as the people in the treatment group. Even so, they are only randomly different, and we know how to characterize random differences with probability statements. Also, because the control does not consist of the same people as in treatment, we cannot estimate an effect for each person, but we can estimate an unbiased average treatment effect by comparing the treatment group average to the control group average.

Causal inference in SCDs can be conceptualized within this potential outcomes model (Pustejovsky, Hedges and Shadish 2014; Shadish, Hedges, Pustejovsky, Rindskopf et al. 2014). The baseline condition (and extrapolation of baseline into the treatment condition) is the source of information about the counterfactual—what the outcome would have been in the absence of treatment, and the treatment condition provides the information about the outcome under treatment. This is also not a perfect measure of the true counterfactual because time passes between baseline and treatment, so that the case could change on its own even without treatment (maturation, regression), or other events other than treatment might have changed the outcome.

However, most SCD researchers do not refer to the potential outcomes model. Instead, SCD researchers approach causation by relying on the ability of high-quality SCDs to rule out most threats to internal validity. The primary tool for doing this is replication, done in a fashion that probes the effect at different predicted time points within and across predicted people on certain predicted outcomes but not others. This rationale is very much like that advanced for any interrupted time series design, but SCDs offer much more opportunity for experimenter control than usual time series designs and so tend to offer stronger inference. Thus, as section 2.1 showed, SCDs are not a unitary design. Even the basic designs are composed of, and can be augmented by, a set of design tactics that SCD researchers use to make the replications complex in a way that can make threats to internal validity unlikely (Sidman 1960).

2.2.1 Ruling Out Threats to Validity

The notion of ruling out threats to internal validity was popularized in a theory of causal inference that is prevalent in psychology and education (Campbell 1957; Campbell and Stanley 1966; Cook and Campbell 1979; Shadish, Cook, and Campbell 2002). Internal validity concerns the inference that an intervention caused an observed outcome. A threat to that inference is any plausible alternative explanation for the observed outcome. Random assignment and its many variations reduce the plausibility of those alternative explanations by distributing them equally (on expectation) over experimental conditions.[6] SCDs reduce their plausibility differently, relying on repeated demonstrations of experimental control over the outcome to make alternative explanations increasingly unlikely to be plausible. For instance, in a multiple baseline design, the researcher can demonstrate that an outcome changes predictably when the intervention is introduced and also changes predictably at different times for each case where those times coincide with the different experimenter-manipulated timing of the introduction of treatment over cases. Finding a plausible alternative explanation for those changes is hard because (a) the alternative explanation would have to work exactly the same way as treatment works on outcome in the absence of treatment, (b) it would also have to occur at different times for each person, and (c) those times would have to coincide with the manipulated time of intervention rather than at some other time. An alternative explanation becomes even less plausible when the design includes multiple demonstrations of an

[6] The exception occurs when random assignment is compromised, such as via differential attrition.

effect, as in an ABAB design, because then it also has to explain how the alternative just happened to coincide with the manipulated time of treatment three times (in this example) within each case.

2.2.2 Make Your Causal Predictions Elaborate

The SCD approach to causation is not new or unique to the SCD literature. The causal logic discussed in this section was anticipated by Sir Ronald Fisher decades ago. Statistician Paul Rosenbaum (2002, p. 327) summarized an exchange between Fisher and the statistician William Cochran, who had a long-standing interest in causal inference when randomized experiments could not be used. Cochran (1965) said: "About 20 years ago, when asked in a meeting what can be done in observational studies to clarify the step from association to causation, Sir Ronald Fisher replied: 'Make your theories elaborate.'" Cochrane went on to clarify: "The reply puzzled me at first, since by Occam's razor, the advice usually given is to make theories as simple as is consistent with known data. What Sir Ronald meant, as subsequent discussion showed, was that when constructing a causal hypothesis one should envisage as many different consequences of its truth as possible, and plan observational studies to discover whether each of these consequences is found to hold."[7]

SCDs do far more than simply predict that the case will respond to the treatment, which would not be an elaborate prediction at all. Rather, they make a set of inter-related causal predictions, for example, that the outcome in a multiple baseline design should change in the predicted direction (and not in the opposite direction) at Time t for Case 1 (that is, a time after the treatment is introduced), Time $t + n_1$ for Case 2 (that is, n_1 sessions after treatment is introduced for Case 1), and Time $t + n_2$ for Case 3 (that is, n_2 sessions after treatment is introduced for Case 1, where $n_2 > n_1$), where all the time points are controlled by the researcher. Similarly, in a reversal design with four phases (ABAB) and multiple cases, the prediction is that the outcome will change in the predicted direction at Time t (a time after the treatment is introduced) for all cases, then will change back to the level of baseline observations at Time $t + (n_1 + 1)$ when the treatment is removed (where n_1 is the number of observations during the treatment phase B_1) and then will again improve in the

[7] Given this quotation is an anecdotal report of an unrecorded speech, we cannot know exactly what Fisher would have counted as being elaborate, and we can think of studies that make more elaborate predictions than most SCDs (e.g., Reynolds and West 1987; Rosenbaum 2002). That being said, SCDs do somewhat more than simply predict that the case will respond to the treatment, which would not be an elaborate prediction at all, and which is the general prediction made by most between-groups nonequivalent control group designs.

predicted direction at Time $t + (n_2 + 1)$ (where n_2 is the total number of time points through the first three phases). Highly cited and thorough books on designing SCDs (e.g., Barlow et al. 2009; Gast and Ledford 2014; Kazdin 2011) show many more ways for SCD researchers to make such elaborate predictions, for example, by concurrently measuring a variable that is predicted not to change during the study or by combining features of the four basic designs in section 2.1 to make more complex predictions. The more elaborate the prediction, the fewer alternative explanations are plausible when the data support the predictions.

2.2.3 Causal Inference and Professional Standards

Methodological standards for SCD studies have achieved considerable professional consensus across education and other disciplines. Perhaps the most widely known example is the U.S. Department of Education's WWC Pilot Standards for single-case designs (Kratochwill et al. 2010, 2013); but many other examples exist (American Speech-Language-Hearing Association, 2004; Council for Exceptional Children (CEC), 2014; Cook et al. 2014; Hitchcock et al. 2014; Horner et al. 2005; Kratochwill and Stoiber 2002; Reichow, Volkmar, and Cicchetti 2008; Tate, Perdices, Rosenkoetter, Wakima, Godbee, Togher, and McDonald 2013). Those standards are based largely on consensus among SCD researchers about what constitutes best practice in SCD research. To briefly summarize the main points of those standards:

1. The intervention must be systematically manipulated by the researcher, not passively observed.
2. The dependent variable must be measured repeatedly over a series of assessment points[8] and demonstrate high reliability. This may be assessed via a psychometric reliability coefficient or more commonly by high inter-rater agreement between two independent assessors or observers of behavior for at least 20 percent of the sessions for which an outcome is generated.
3. The design must demonstrate an intervention effect at three different points in time (e.g., with at least three separate series in a multiple baseline or through at least three manipulations of the independent variable in an ABAB design).
4. In general, each phase of the design (i.e., baseline, intervention) must contain at least three to five data points that document a pattern of performance. The exception to this is alternating

[8] The Pilot Standards discuss the number of observations per phase, typically five, but the excerpts are sufficiently complex that we do not provide details here.

treatment designs where conditions are intended to alternate rapidly, often with only one outcome observation per alternation. The WWC Pilot Standards require such designs to have a minimum of five AB alternations (or BC alternations if two treatments are alternated) so that the outcome under A is observed five times and the outcome under B is observed the same number of times.

5. The studies must meet the 5-3-20 rule: (1) at least 5 SCD studies are available that each meet evidence standards; (2) those studies are conducted by at least 3 different research teams with no overlapping authorship; and (3) the combined number of cases across all qualified studies is at least 20.

The third standard is the key one for elaborating causal hypotheses, predicting not just that a treatment will work, but also that it will work with different people at different predicted times or within the same person at three different predicted times. That standard could be met by a reversal design (e.g., ABAB) that starts with baseline observations in the absence of treatment, then introduces treatment, and then removes and reintroduces the treatment again. Or it could be met by a multiple baseline design in which each of at least three cases only has one baseline and one intervention phase but the introduction of the intervention is staggered over time for multiple cases.

In addition to the WWC, several other groups have forwarded similar standards for SCDs that are acceptable for inclusion in evidence-based practice reviews (see Smith 2012 and Wendt and Miller 2013, for review and comparison). These include the American Psychological Association (APA) Division 16 Task Force on Evidence-Based Interventions in School Psychology (Kratochwill et al. 2010; Kratochwill and Stoiber 2002), the Council for Exceptional Children (CEC) [2014; Cook et al. 2014; Horner et al., 2005], the American Speech-Language-Hearing Association (ASHA 2004) , and the Oxford Centre for Evidence-Based Medicine (Howick et al. 2011).. Such groups often differ from WWC in many ways, most notably that they are not often funded by government agencies but rather consist of researchers with shared substantive interests is areas such school psychology, exceptional children, or speech and hearing disorders. The recommendations from these groups regarding good SCDs are quite similar, however. Of course, educational researchers are not the only ones interested in standards for SCDs. In medicine, several groups have developed or published standards for reporting SCDs (called N-of-1 trials in medicine) (Shamseer et al. in press; Tate et al. in preparation; Vohra et al. in press). So standards for SCD research are not defined solely by

educational researchers, though little work has been done to communicate across disciplines, these three medical examples being the only ones we know.

2.3 Single-Case Designs: Some Salient Evaluative Issues

No experimental design is perfect. The randomized experiment, for example, can suffer from differential attrition of cases from conditions, and that can bias results. SCDs have their limitations, as well. This section raises a few salient issues that may arise when evaluating SCDs, especially relative to other experimental designs. A thorough analysis of such issues is beyond the scope of this paper, but this section can begin that discussion.

2.3.1 Two Reasonable Concerns

Statistical Characteristics of Effect Estimates. Randomization guarantees[9] that the standardized mean difference parameter δ is a causal effect. Technical properties of the estimate (e.g., small sample bias correction) guarantee that the estimate g is an unbiased estimate of δ. Designs that do not use random assignment do not have these properties, e.g., it would be possible for them to yield an unbiased estimate of an effect size parameter that is a biased estimate of the causal effect. This is pertinent because the claim in this paper that certain SCD estimators yield results that are in a comparable metric as in between-groups experiments should not be misconstrued as a claim that those estimators also have all the other properties of an estimator from a randomized experiment. Clearly, then, SCD research would benefit from further study of the statistical characteristics of its effect estimators, especially about what population parameter is being estimated both on its own terms and in relation to that estimated by the randomized experiment.

Also worth further development is which causal estimators are being estimated in all these designs. For example, most randomized experiments choose to estimate an average treatment effect (ATE), leaving the effect of a treatment on an individual an open question. SCDs in principle allow such statements about individuals. However, much work is still needed to understand how concepts like ATE, intent-to-treat (ITT) and effect of treatment on the treated (TOT) apply to SCD effect estimates.

[9] Randomization guarantees this on expectation (on average), not for any individual study.

Applicability Across a Range of Causal Questions. SCDs typically involve intensive observation of and interaction with a single case, including gathering many consecutive observations over time and constantly monitoring the responsiveness of the outcome to conditions over time. The resource requirements to do this can exceed what is possible for many questions about the effectiveness of interventions. For example, policy questions about the effects of a nationally implemented program (e.g., Head Start) are unlikely to be amenable to study using SCD methodology. Therefore, both SCD researchers and between-groups researchers would benefit from a more systematic understanding of the kinds of problems, interventions, settings, and outcomes that are more or less amenable to the use of SCD methods.

2.3.2 Concerns That May Be Less Compelling

Generalizability. In part because of the label SCD (or N-of-1 trial)[10], some researchers raise concerns about the generalization of results from SCDs. Usually the implicit assumptions are twofold: (1) that generalizations are mostly facilitated by large sample size so that single cases are inherently disadvantaged and (2) that the main target of generalization is some population represented by the case, that is, we want to know about generalizations to other cases. In some respects, the concern about generalization has merit. All other things being equal, the replication of SCD results over an increasingly large number of cases would indeed have both inferential and statistical benefits. SCD researchers themselves often conclude their articles with such a call. That being said, the two implicit assumptions of this concern are overstated.

Regarding the first assumption, a review of how scientists (including those who do between-groups experiments) actually make generalizations in their work (Shadish, Cook and Campbell 2002, Chapter 11) suggests that they rarely rely much on any systematic sampling rationales or methods (with a few obvious exceptions such as survey research). Much more often, they use appeals to apparent similarity of study operations to the target of generalization, identification of the limits of generalization through the study of moderator variables, making interpolations and extrapolations from what is observed to what is not observed, and generating explanatory models about the

[10] The label may mislead some readers to think that most such studies have only one case, whereas SCD studies have a median of three or more cases (Shadish and Sullivan 2011).

mechanisms underlying an effect that help predict the conditions under which the effect could be replicated. Most salient to the present point, increasing sample size is generally not a method by which scientists improve their generalizations (though they may increase sample size for other reasons such as improving power).

Regarding the second assumption, scientists actually want to generalize not only about cases but also about how the effects of a treatment might vary across settings, across different ways in which the treatment could be implemented (e.g., increasing dose, kinds of providers), and across observations such as variations in how the outcome is conceived and measured (e.g., if results of an experiment showing a treatment improves the prostate-specific-antigen test result in prostate cancer generalize to the treatment's effectiveness for survival) (cf., Cook 1990; Cronbach 1982; Shadish, Cook, and Campbell 2002). Sampling strategies, whether formal like random sampling or informal like choosing the most representative instance, are frequently impractical or sometimes logically irrelevant to such generalizations (e.g., it is rare to have a list of all possible treatments from which to sample). Instead, scientists use other principles to help them make generalizations. Shadish, Cook, and Campbell (2002), for example, reported results of a review of how scientists generalize in many different contexts, from instances to categories, from measures to constructs, from diverse findings in a literature to scientific principles, from animal models of human psychopathology to human symptoms, from animal models of toxic chemicals to their effects on humans, from epidemiological studies of second hand smoking to causal inferences about smoking and cancer, from animal testing of drugs to conclusions about drug effects in humans, and from laboratory studies of psychotherapy efficacy to conclusions about its effectiveness in therapy clinics. That review distilled five principles of causal generalization that scientists routinely use in addition to sampling methods:

- *Principle 1: Surface Similarity*: Scientists generalize by judging the apparent similarities between the things that they studied and the targets of generalization—for example, studies of the effects of secondhand smoke in the workplace seems more similar to the public settings at issue in policy debates than do studies of smoking in private residence, and animals with wings seem more similar to our prototypical understanding of birds than do animals without wings. This is the most basic principle.
- *Principle 2: Ruling Out Irrelevancies*: Scientists generalize by identifying those attributes of persons, settings, treatments and outcome measures that are irrelevant because they do not change a

generalization—for example, that size is irrelevant to membership in the category of bird, that length of follow-up is irrelevant to the effect of psychotherapy, or that the location of a cognitive science research lab is irrelevant to the finding that people tend to use groups of seven to remember things.

- *Principle 3: Making Discriminations*: Scientists generalize by making discriminations that limit generalization—for example, that child psychotherapy works in the lab but might not in the clinic, that a claim to have a new measure of a trait involves discriminating that trait from other ones, or that any animal with both feathers and wings falls within the boundaries of the category of birds but all other animals fall outside that category.
- *Principle 4: Interpolation and Extrapolation*: Scientists generalize by interpolating to unsampled values within the range of the sampled persons, settings, treatments and outcomes, and much more difficult, by extrapolating beyond the sampled range—for example, that the effects of cytotoxic chemotherapies for cancer increase as the dose increases until the point where they would cause the patient to die, or that effects of toxic chemicals on small mammals will generalize to much larger and more biologically complex humans.
- *Principle 5: Causal Explanation*: Scientists generalize by developing and testing explanatory theories about the target of generalization—for example, that the construct called "effects of stimulants" includes both scratching in primates and rearing in rats because the biological mechanisms underlying both these behaviors are the same.

These principles of causal generalization are usually just as applicable to SCDs as they are to between-groups studies (Shadish 1995). Consequently, we see little reason to think that results from any individual SCD study will encounter more serious generalization problems than will results generated by other experimental methods.

Matters Already or Easily Addressed in SCD Work. Sometimes concerns reflect communication gaps between communities with different specialties. These might include blinding, carryover effects, dependence on time intervals, inter-rater agreement, and terminology not shared across research communities. Each of these either has already been addressed in the SCD community or can be addressed without major obstacles. Regarding blinding, for example, blinding of outcome raters to conditions and cases is both common and recommended best practice in SCD research. Carryover effects are plausible, especially when treatments have lasting effects, but variants such as the

multiple-baseline design are intended precisely to address such concerns. Concern about the extent to which results might depend on how time intervals are defined (e.g., by session or by chronological time) are entirely legitimate but easily remedied (asking SCD researchers to report chronological time is a recommendation of this report). Ensuring high inter-rater reliability has long been a central tenet of SCD research, and nearly all studies report high reliability levels. Terminological differences are nontrivial (e.g., not all between-groups researchers immediately recognize the term "partial interval recording"), but communication among research communities will do much to remedy this problem.

3. Effect Sizes and Single-Case Designs

As documented in section 2.2.3, SCD studies can be included in evidence-based practice reviews (e.g., U.S. Department of Education 2014a), whether those reviews are conducted by WWC or by independent researchers doing narrative or quantitative reviews on a topic of interest. In fact, however, with a few exceptions, SCDs are often not included. When they are included (e.g., U.S. Department of Education 2014b, c), they are often kept separate from reports of group studies. In one case, an SCD was accompanied by a footnote that "The WWC does not currently calculate effect sizes for single-case design research and does not currently summarize findings across single-case design studies" (U.S. Department of Education 2014c, p. 32). One key reason for that omission is the topic of the present paper: a lack of statistical analysis in general, and of effect sizes in particular, in most SCD reports often leaves typical reviewers and policymakers unsure how to interpret results from SCDs, how to combine results from different SCDs on similar questions, how to compare results from SCDs to results from other experimental methods (even if the two sets of results are not combined), or how to combine SCD results with those from those other methods to produce a single overall effect size. This paper presents an approach to these problems, the inclusion of standardized between-case effect sizes in SCD reports and the reporting of data in such a way that others can easily compute effect sizes.

We are not proposing that calculation of effect size replace the traditional use of visual analysis in single-case research but rather that effect sizes become an additional source of information to help understand the size of an effect. We support the continued practice that single-case studies should be assessed first to determine if the design will allow assessment of a functional relation (e.g., Kratochwill et al. 2010, 2014), second to determine if the data within the design will allow support for the presence of a functional relation (i.e., does the design allow for three demonstrations of an experimental effect), and third to examine the social importance of the demonstrated effect.[11] Even more strongly, it is unwise for the SCD researcher to rely solely on effect sizes to draw conclusions about cause and effect relationships—the logic of causal inference is mostly design dependent, not

[11] SCD researchers have long been concerned with studying outcomes that are socially important, that is, with implications that stakeholders agree will make a useful difference to the case (Gast and Ledford 2014). So demonstrating social validity is an integral part of most SCD reports. For example, social validity might be supported if a teacher or parent would agree the change in the outcome as a result of treatment is a change that makes a meaningful difference to a child's achievements or happiness.

statistics dependent. Interpretation of overall study results will also need to consider failures to replicate over or within cases or large differences in consistency or magnitude of effect over cases. But to all this current practice, we add the recommendation that routine documentation of between-case effect size become an additional part of this practice and that documentation of effect size will be important to subsequent inclusion of SCDs in evidence-based practice reviews.[12]

3.1 What Are Standardized Effect Sizes?

In this paper, we define standardized effect sizes as statistics that describe the magnitude of an effect on a common scale. Such effect sizes might not be needed if all studies testing the effect of a reading intervention used the same measure of reading fluency (e.g., oral reading fluency) to assess student performance. But that is rarely the case in educational research, where studies use a wide variety of outcome measures. For example, studies of interventions to improve early reading competence may use oral reading fluency, standardized state tests, comprehension, or nonsense word fluency as outcome measures. These different outcome measures each have a different range, use different response formats, and have different means and variances that cannot be combined in raw form—just as one could not simply combine height in inches with height in centimeters without first putting them on a common metric. As we use the term in this paper, standardized effect sizes put outcomes on that common metric (e.g., all with a mean of 0 and a standard deviation of 1, or all ranging from 0 to 100 percent). In the group experimental literature, such effect sizes are well developed, with examples that include the standardized mean difference statistic, the odds ratio, and the correlation coefficient (Shadish and Haddock 2009).

3.1.1 How Do Standardized Effect Sizes Help Evidence-Based-Practice Reviewers?

Standardized effect sizes help solve some important problems reviewers face when including SCDs in evidence-based-practice reviews. They allow reviewers to

[12] To anticipate one argument, we recognize that some SCD reports sometimes already include an effect size estimate, typically drawn from the overlap statistics (e.g., Percentage of Nonoverlapping Data; Tau-U), but sometimes regression based. Shortly, we will describe why those effect sizes will not meet the need we describe in this paper. But first, consider in more detail the nature of standardized effect sizes.

- *interpret* results from SCDs using the same conventions they apply to other designs—effect sizes and confidence intervals and associated significance tests (either directly computed or by seeing if the confidence interval excludes zero); and
- *combine* results from different SCDs on the same question, for example, studies of interventions to improve early reading competence that use diverse outcome measures such as oral reading fluency, standardized state tests, comprehension, or nonsense word fluency.

In addition, three standardized effect sizes (Hedges, Pustejovsky, and Shadish 2012, 2013; Pustejovsky, Hedges, and Shadish 2014; Swaminathan, Rogers, and Horner 2014) also allow the reviewer to

- *compare* results from SCDs to results obtained from between-groups studies like randomized experiments; and
- *combine* SCD results with those from other designs, if that seems appropriate.

Effect sizes that allow these latter two features are called between-case effect sizes in this paper, as we explain in the next section, but they are sometimes also called design-comparable effect sizes because they report the effect size in a metric that is the same for many different kinds of designs (Pustejovsky, Hedges, and Shadish 2014; Swaminathan, Rogers, and Horner 2014; Van den Noortgate and Onghena 2008).

3.1.2 Within-Case Versus Between-Case Effect Sizes

Many effect sizes for SCDs exist (Shadish 2014b). They all meet the second criterion above, that is, they allow combining results from different SCDs on the same question. Some also meet the first criterion in that they use well-developed statistical methods like regression so that the standard errors, confidence intervals, and significance tests have known accuracy[13] (e.g., Beretvas and Chung 2008). However, very few SCD effect sizes meet all four criteria. The main reason has to do with the difference between what we will call for pedagogical purposes *within-case* versus *between-case* effect sizes. This section aims to introduce the concept of within- versus between-case effect sizes but not

[13] We are not saying the other effect sizes have inaccurate standard errors, confidence intervals, or significance tests, only that their accuracy is not clear given their lack of formal statistical development.

their detailed computation. Those computations will differ, sometimes substantially, depending on exactly which specific version of a within- or between-case effect size is chosen, and the computations are not illustrated here. Shadish (2014b) provides more detail about the specifics of some of the current within- and between-case effect sizes for SCD research, and this paper contains *worked-through examples* that provide access to more computational details.

The vast majority of previous SCD effect sizes are within-case effect sizes; this includes both regression-based and overlap statistics.[14] They might, for example, subtract baseline mean outcomes ($M_{Baseline}$) from treatment mean outcomes ($M_{Treatment}$), and then divide that mean difference by within-case variability (s_{within},) in outcome observations. Say a case with an ABAB design has 20 observations over time, 8 in each of the 2 baseline (A) phases and 12 in each of the 2 treatment (B) phases. The average of the 16 baseline observations gives $M_{Baseline}$, the average of the 24 treatment observations gives $M_{Treatment}$, and the square root of the variance of the 40 observations (calculated within each phase and pooled) gives s_{within}, which is a measure of how much one case's observations tend to differ from each other within phases (how variable are one case's observations).

Beretvas and Chung (2008) took this kind of approach when they used multiple regression to create a *d*-statistic for each case. The result was a *d*-statistic that converted any outcome variable into a common metric with well-understood statistical properties. This metric can be used (either meta-analytically or narratively) to compare the outcomes from one case to another case or the outcomes of cases in one SCD study to outcomes in another SCD study. However, it cannot be used to compare results from SCD studies to results from between-groups studies, which is often a desirable goal in narrative and meta-analytic evidence-based-practice reviews. To explain why, we first need to define a between-case effect size.

Only a few SCD effect sizes are between-case effect sizes (Hedges, Pustejovsky, and Shadish 2012, 2013; Pustejovsky, Hedges, and Shadish 2014; Swaminathan, Rogers, and Horner 2014). In between-groups research, such effect sizes typically compare mean outcomes from the treatment group to those from a comparison group but then standardize that mean difference with the variability

[14] If an SCD effect size is not explicitly mentioned below as being a between-case effect size or is not explicitly derived from them or similarly to them, then the reader can assume the effect size is probably a within-case effect size.

between cases (how much the observations differ over cases within the same group). The usual standardized mean difference (d) statistic computed for a randomized experiment is:

$$d_{between} = \frac{M_{Treatment} - M_{Comparison}}{s_{between}} \quad (1)$$

So if a randomized experiment had 20 people in the treatment group and 20 in the comparison, the average of the 20 treatment case outcomes gives $M_{Treatment}$, the average of the 20 comparison group outcomes gives $M_{Comparison}$, and $s_{between}$ is computed as the square root of the average of the variance of the observations in the 20 treatment group cases and the variance of the observations of the 20 comparison group observations (usually an average that weights the variances by a function of sample size so that the variance from a group with a larger sample size would get more weight).[15]

The three existing between-case effect sizes for SCDs (Hedges, Pustejovsky, and Shadish 2012, 2013; Pustejovsky, Hedges, and Shadish 2014; Swaminathan, Rogers, and Horner 2014), which will be explained further in Section 2.2 of this report, aim to estimate $d_{between}$. The computational details differ over the three methods and in any case are too complex to present here. At a general level, however, these equations highlight two differences between $d_{between}$ and d_{within}. The first is that s_{within} can be computed with data from just one case. In contrast, $s_{between}$ requires data from multiple cases—you cannot compute the extent to which cases vary from each other unless you have multiple cases. The second is that the means in the numerators of both equations are slightly different conceptually. For $d_{between}$, the numerator takes into account not only how treatment changes outcomes within a case compared to no treatment but also how this mean difference occurs across cases.

As the above discussion suggests, the key reason why these two statistics, d_{within} and $d_{between}$, are not comparable to each other is that they standardize differently. In general, how one person differs within himself or herself over time (s_{within}) is often quite different from how several people differ from each other at the same time ($s_{between}$). Take, for example, the weights of people. Imagine a person records his or her weight each day for 20 days. Typically, the variance of those weights is likely to be small—our daily weight does not go up or down much over 20 days. Then imagine taking the weight of 20 different people at one time. Weight is likely to vary substantially over people. Theory would

[15] This paragraph describes the variance that most researchers are quite familiar with. Later, this paper will introduce a different kind of variance, the variance of a sample estimate of the effect size (see section 6.2.1, equation 2), that tells how a set of effect sizes that differ only in their sample will differ from the population effect size by chance.

suggest that $s_{between}$ will be larger than s_{within} because between cases variation logically includes within person variation.[16] Of course, how much bigger $s_{between}$ will be than s_{within} will depend on what variable we record; but experience suggests $s_{between}$ will usually be considerably larger than s_{within}. The point is that d_{within} and $d_{between}$, are simply not comparable to each other and so cannot be compared or combined. [17]

So an evidence-based practice review that includes (1) SCDs reporting d_{within} and (2) between-groups studies reporting $d_{between}$ cannot, for example, say that results from the SCDs are bigger or smaller than results from the between-groups studies, and it cannot combine the two. In fact, because we often have no idea what the relationship is between d_{within} and $d_{between}$, reviewers cannot say much at all about how outcomes from SCDs compare to those from between-groups experiments. This risks discouraging reviewers from including SCDs in evidence-based practice reviews or, even worse, risks the reviewer not understanding this lack of comparability and making the comparison anyway. *For all these reasons, then, we assume in this paper that between-case effect sizes are most desirable if the purpose is to have SCDs included in evidence-based practice reviews that also review between-groups methods.*

This is not to discourage the use of within-case effect sizes for other purposes. Within-case effect sizes allow researchers to draw conclusions about each case separately, something between-case effect sizes cannot currently do. They also have great credibility because SCD researchers are trained to examine features of outcome data such as the variability of outcome observations in baseline versus treatment conditions within a single case. Within-case effect sizes can be useful when the review will include only SCD studies, and the review will not draw any comparisons to results from between-groups studies (though better statistical development of some of the within-case effect sizes would still be useful). Nonetheless, the usefulness of within-case effect sizes does not extend to doing evidence-based practice reviews that include between-groups designs, and we presume the latter is both common and a highly desirable goal if SCD research is to take its proper place in the evidence-based practice literature as a whole. To speculate, there are scientific reasons to think that effect sizes in SCDs should be larger than those in most between-groups studies. The treatments in

[16] This fact may not be obvious without closer study, but is true (see Hedges et al. 2012, 2013).

[17] Note that differences in the way individuals are sampled in different SCD studies may contribute to heterogeneity in effect sizes computed from SCDs, just as differences in sampling plans of between-subjects designs (BSDs) contribute to heterogeneity of effect sizes. Differences in sampling plans therefore should be considered as potential explanations of between-study heterogeneity in meta-analyses of SCDs, just as they are in BSDs

SCDs are targeted to specific individuals and their specific problems. If the targeting is effective, the treatments should make a difference for more individuals than in between-groups designs where the treatments are (usually) not specifically targeted, and the treatment effect can sometimes be conceptualized in terms of the "fraction of the subjects that responded to treatment." If this proves to be the case, it would have implications for whether or not to pool results from SCDs and between-groups studies, which is the topic of section 6.3 of this paper.

3.1.3 Serial Dependence in SCDs and Its Effect on Effect Sizes

Observations (or more technically, their errors) taken over time from the same case may not be independent of each other. Most statistics, including effect sizes, assume independence of errors, so we cannot use those statistics easily to compute effect sizes for SCDs. Failing to take serial dependence (sometimes called autocorrelation) into account will result in the variance in the denominator of the effect size to be larger or smaller than without autcorrelation, depending on whether the serial correlation is below or above zero, respectively. To obtain an effect size in the same metric as the usual between-groups effect sizes, where the pertinent autocorrelation is legitimately presumed to be zero because the within groups variance are based on independent group members, adjustment for autocorrelation is required.. Some current effect size estimators adjust for serial dependence, and some do not; we talk about them more shortly.[18]

3.2 Worked-Through Examples and Related Application Issues

This paper presents many examples of between-case effect sizes and analyses and meta-analyses of those effect sizes. However, as we said previously, the paper does so at a conceptual level, not at a level that shows all computations. The reasons are twofold. First, computational details that include relevant prose, equations, and annotated syntax, are quite long, easily requiring at least a doubling of the length of an already long paper. Second, those details already exist in either the published literature or in electronic format on accessible webpages. Instead of providing those details, this paper cites pertinent references that themselves contain the worked-through examples and includes

[18] Assessing the extent to which autocorrelation is a problem can be difficult because, for example, it can be affected by trend so that effect sizes that adjust for trend may also reduce the problem of autocorrelation. However, we have too little knowledge about this to make strong statements. Autocorrelation is probably sufficiently prevalent (e.g., Shadish and Sullivan 2011) that simply ignoring it may not be a good decision.

call-out boxes providing sufficient detail to allow the researcher to locate those examples for further use. Search this manuscript for the term *worked-through* to find these examples. Five references that are good starting points for conducting the analyses described in this paper are Marso and Shadish (2014), Shadish, Hedges, and Pustejovsky (2014), Shadish, Hedges, Pustejovsky, Boyajian, et al. (2014), Pustejovsky, Hedges, and Shadish (2014), and Swaminathan, Rogers, and Horner (2014). In addition, those call-out boxes provide hyperlinks to websites that provide considerably more detail, though that is not possible in all cases due to copyright law.

3.2.1 Software for Computing Effect Sizes and for Doing Meta-Analysis

An important distinction in these examples is between computing an effect size versus doing a meta-analysis of multiple effect sizes. The computation of an effect size for SCDs does require software (e.g., Shadish, Hedges, and Pustejovsky's 2014 SPSS macro[19]) specific to the effect size of interest. However, the researcher can use any program capable of doing meta-analysis to combine and analyze multiple effect sizes, whether from SCDs or not. SPSS has limited capacity to do state-of-the-art meta-analysis (Field and Gillett 2010; Wilson 2006). Both SAS (Arthur, Bennett, and Huffcutt 2001) and especially Stata (Sterne 2009) provide many excellent routines (see Shadish, Hedges, Pustejovsky, Boyajian, et al. 2013 for a meta-analysis of SCD data using Stata). The freestanding program Comprehensive Meta-Analysis (Borenstein, Hedges, Higgins, and Rothstein 2005, 2009) has a point-and-click interface and incorporates many major features of modern meta-analytic statistics, but it is more expensive. The metafor package (Viechtbauer 2010) in the free computer program R (R Development Core Team 2012) has extensive documentation complete with examples and syntax for doing a wide array of analyses and producing publication-quality graphics (see http://www.metafor-project.org).

3.2.2 Computing Effect Sizes Proactively and Retroactively

Implicit in the examples in this paper is also the fact that one can apply these methods either to new studies or to studies that were previously published in the literature. For a new study, the researcher presumably has the original raw data. For past studies, the researcher will usually have to digitize the

[19] The macro may be downloaded from http://faculty.ucmerced.edu/wshadish/software/software-meta-analysis-single-case-design.

data from published graphs (see section 4.4). In either case, the dataset will need to be in specific formats that vary from software to software and that are referenced in the working examples that are cited throughout. However, that does not make this process any different from any other statistical analysis given that different software packages nearly always have slightly different input format requirements.

3.3 Between-Case Effect Sizes for SCDs

As noted previously, three between-case effect sizes exist for SCDs (Hedges, Pustejovsky, and Shadish 2012, 2013; Pustejovsky et al. 2014; Swaminathan, Rogers, and Horner 2014).[20] The details of each effect size differ, and some are far more accessible than others to the SCD researcher. Here we provide a brief overview of these three effect sizes, in alphabetical order by author.

Hedges and colleagues (2012, 2013) created two closely related between-case *d*-statistics for SCDs, one for multiple baseline designs across individuals and one for reversal (e.g., ABAB) designs. Both *d*'s estimate the same parameter, take into account autocorrelation, and are adjusted for small sample size bias. The authors developed power analyses for these statistics to predict the number of cases and time points needed to reliably detect an anticipated effect size in the planning of a study, given assumptions about autocorrelation and between-case versus total variance. The effect size assumes normality of the residuals about phase means within people,[21] no trend (or that the time series had been detrended, an option in the software), and that the treatment effect is constant over cases. Consider the last assumption a bit more. Within one case, the treatment effect is measured as the change in mean outcome between phases, so an ABAB design gives three estimates of the treatment effect in a case; those three estimates do not have to be the same, and they are averaged in the effect size. The assumption, though, is that this average effect for each case is the same for all cases (within sampling error). Hedges and colleagues (2012, 2013) showed how to test several of these assumptions and, in some cases, how to proceed with the analysis if the assumptions are violated. The method requires three cases within a study in order to estimate all the pertinent statistics. An

[20] Other authors have described possible directions such work could take without fully specifying relevant models or software to implement them (e.g., Van den Noortgate and Onghena 2008).

[21] This is often described as normality of observations themselves, but the assumption is really about normality of residuals. Residuals could be normally distributed even when the outcome metric itself is not (e.g., a count). Further, normality of the outcome variable does not guarantee normality of residuals.

SPSS macro both for estimating the effect size and for doing power analyses is free for download, with input either through syntax or a graphical user interface, and with a manual to show the process.

Worked-Through Examples for Hedges et al. (2012, 2013).

The webpage http://faculty.ucmerced.edu/wshadish/software/software-meta-analysis-single-case-design has an SPSS macro for computing this effect size and its power for a single study. It also allows downloading of an extensive manual showing how to use the macro. This includes how to install the macro, provide input data, execute the macro using either syntax or a graphical user interface, and interpret results of the macro. The manual includes snapshots of input and output at all phases of the process. Shadish, Hedges, and Pustejovsky (2014) provide similar detail on effect size computation and power for single studies, using worked-out examples, with the raw data in an appendix so the user can replicate the results.

Pustejovsky et al. (2014) extended the work of Hedges and colleagues (2012, 2013), conceptualizing it in the more general framework of multilevel modeling. All the characteristics and assumptions of this extension remain the same as the original (e.g., it requires a minimum of three cases), but it also allows linear time trend, an interaction between trend and treatment for the case of multiple baseline designs across individuals and heterogeneous coefficients for all these effects over cases. That is, trend, treatment effects, and their interaction can vary over cases by more than would be expected by chance (sometimes called a random effect; see the last paragraph of section 6.2.1). The authors show that this more flexible model requires larger sample sizes to estimate well, particularly an increase in the number of cases as high as 12 to accurately estimate random effects—though these 12 cases can come from more than one study on the same question that are all included in one review.[22] Of course, one need not estimate all these fixed and random effects; models that are more constrained (e.g., no random treatment or trend effects) require fewer cases (e.g., six to nine). For instance, the Hedges et al. (2012, 2013) estimators would usually have more statistical power because

[22] The outcome and measurement protocol should be the same for all cases across studies to include all cases in the same effect size. When the specific protocols differ over studies, it may be better to compute effect sizes separately for each study and then aggregate them using meta-analytic procedures.

they make more assumptions (e.g., assuming no trend, no random effects). The old adage holds, you need either more assumptions or more data for the same design sensitivity—whether using SCD, randomized experiments, or any other design. All these estimates of sensitivity, of course, will vary greatly depending on the size of the effect, which we know little about at this early stage of development (more on that later). The authors have an R package (R Development Core Team 2012) allowing estimation of various versions of this effect size.

Worked-Through Examples for Pustejovsky et al. (2014).

The webpage http://blogs.edb.utexas.edu/pusto/software/ has instructions for downloading and installing an R package called *scdhlm* that implements all the analyses in Pustejovsky et al. (2014). The original article shows the results (but not the syntax) for one example of how to compute an effect size for one study, with and without taking trend into account. However the authors have written a vignette (R terminology for an extended example of usage) that provides instruction in how to use the effect size estimation functions corresponding to the Hedges et al. (2012, 2013) papers and the Pustejovsky et al. (2014) paper. To see the vignette, run the following lines from the R command prompt:

```
install.packages("devtools")
library(devtools)
setInternet2(use = TRUE)
download.file("https://utexas.box.com/shared/static/9tikotwuvcsam8hjssuu7fkp5o82kumt.zip",
destfile = "scdhlm_0.2.1.zip")
install.packages("scdhlm_0.2.1.zip", repo = NULL)
library(scdhlm)
vignette("Estimating-effect-sizes")
```

The vignette shows how to download several different data sets and compute variations of the between-case effect sizes that Pustejovsky et al. (2014) developed.

Swaminathan et al. (2014) developed a Bayesian multilevel approach allowing trend and interactions, assuming normality of the residuals about phase means within people, with all coefficients allowed

to vary over cases. The most noteworthy advance of the Swaminathan et al. (2014) approach over both the other two approaches is that Bayesian methods can allow more stable parameter estimation, especially regarding variability and random effects, in the smaller sample sizes often typical of SCD research. Their between-subjects *d*-statistic seems to well approximate the results from Hedges et al. (2012, 2013). It also requires a minimum of three cases.

Worked-Through Examples for Swaminathan et al. (2014).

Swaminathan et al. (2014) give an example of their Bayesian analysis, and appendix A of that publication provides annotated syntax and input data for the free OpenBUGS computer program (Lunn, Spiegelhalter, Thomas, and Best 2009). The authors are currently revising their SINGSUB computer program (Rogers and Swaminathan 2009) to incorporate these analyses. Interested researchers can use the syntax in that appendix or they can contact the authors at Swami@uconn.edu for updated software.

3.3.1 Evaluating Between-Case Effect Sizes for SCDs

A good between-case effect size for SCDs should account for trend and dependencies of observations within cases; should show how to do a power analysis; should use appropriate statistical distributions for such diverse metrics as continuous variables, counts, and rates; and should have a user-friendly interface that SCD graduate students and applied researchers can easily use (Shadish 2014b). None of the above between-case effect sizes do all these things. We discuss below how each of the three between-case effect sizes perform on these key evaluative dimensions. Table 1 summarizes the discussion, with a check mark (✓) indicating that the effect size either does or can accommodate the criterion reasonably well. Table 1 necessarily oversimplifies the more nuanced discussion that follows. The table awards check marks conservatively, that is, only when it is very clear the effect size as currently developed meets the criterion; it seems likely that these methods will improve rapidly to address gaps identified in table 1.

Table 1. An evaluative summary of three between-case effect sizes

Criterion	Hedges	Pustejovsky	Swaminathan
Trend		✓	✓
Dependency	✓	✓	✓
Power Analysis	✓		
Non-normality Accessibility	✓		

Trend. The Pustejovsky et al. (2014) and Swaminathan et al. (2014) estimators allow linear trend and, in theory, could incorporate nonlinear trends but rarely do. The Hedges et al. (2012, 2013) estimator assumes no trend. Instead, those authors show how to test for trend and offer an option in the software to detrend the data with linear or nonlinear terms prior to computing *d*. However, the effects of detrending are unclear. Some evidence suggests that detrending results in smaller effect size values than might otherwise be the case using Gorsuch's (1983) differencing and trend analysis (Brossart, Parker, Olson, and Mahadevan 2006; Manolov, Arnau, Solanas, and Bono 2010), but estimating baseline trend and controlling for it as in Allison and Gorman's (1993) model might not lead to conservative results (Brossart et al. 2006; Campbell 2004; Manolov and Solanas 2008). So, detrending needs further study.

Dependency. Previously we noted that observations within a case (or more technically, errors of observations) cannot be presumed to be independent but, rather, are serially correlated or autocorrelated so that one observation is related to one or more of the observations that came before it. Failure to adjust for this can lead to biased estimates and standard errors. Fortunately, all three methods can account for the fact that observations within cases cannot be assumed to be independent. They do so either by estimating the autocorrelation as part of effect size estimation (or estimating a closely related autoregressive model) or by using random effects. Either approach seems to result in reasonable standard errors (Gurka, Edwards, and Muller 2011).

Power Analysis. Hedges et al. (2012, 2013) have associated power analyses to help estimate the number of cases, and observations within cases, that are needed to detect an anticipated effect size (Shadish, Hedges, Pustejovsky, Boyajian et al. 2014; Shadish, Hedges, Pustejovsky, Rindskopf et al.

2014). The other two estimators (Pustejovsky et al. 2014; Swaminathan et al. 2014) have not shown how to do power analyses. Shadish and Zuur (2014) have compared power for some related multilevel models, however, and show that power is lower for them because they try to estimate more characteristics of the data (e.g., autocorrelation, trend, random effects) with the same amount of data. As noted earlier, the tradeoff is a common one in research—one can increase power by making more assumptions (like no trend or fixed effects in Hedges et al. 2012, 2013) or by gathering more data when estimating more complex models (Pustejovsky et al. 2014; Swaminathan et al. 2014). More data are preferable but not always possible or practical.

Some researchers are skeptical that power analysis contributes much to SCD studies, often claiming that increasing sample size is the only way to improve power but is nearly always infeasible. The claim is wrong on two counts, that other ways exist to improve power, and that while increasing sample size is certainly not possible in some instances, it is possible in many others. Section 5.1 considers this topic in more detail.

Non-normality. All three approaches assume normality of the residuals about phase means within people. SCD researchers may wonder about whether this will be true of the count and rate data so ubiquitous in SCD research. Although more research on this is certainly needed, assumptions about normality of residuals may not be as large a concern as one might think. One reason is that assuming normality of the residuals about phase means within people is not the same as assuming normally distributed raw data. Normally distributed residuals may be plausible even when the original raw data are counts or rates. Only inspection of the residuals will reveal that, e.g., a plot of Pearson residuals over time for each case. Researchers may also wonder about situations in which there are very low frequencies in some (e.g., baseline) phases, such as a high frequency of zeros (sometimes called zero inflation in the statistical literature). The use of an effect size using between-case variation often lessens (but may not eliminate) this concern because the denominator includes variability both within and between cases so that zero inflation in any particular phase may have less impact.

Second, some reason exists to think the normality assumption may not always be crucial for the accuracy of the results. For example, Shadish, Hedges, and Pustejovsky (2014) note that "the *g* statistic may be considered robust to some violation of the normality assumption, because both the numerator and denominator of the effect size are precisely unbiased even when the data are not

normally distributed" (p. 131). Presumably this extends to the Pustejovsky et al. (2014) estimator given how closely related it is to the Hedges et al. (2012, 2013) estimator, and it probably extends to the Swaminathan et al. (2014) estimator as well. However, effect sizes specifically designed for count and rate data still need development and need more extended study to see whether using other distributional assumptions makes substantial difference to the answers. Tentatively, then, we recommend using existing between-case effect sizes despite this potential problem given that doing so may not pose a serious problem to the results.

Accessibility. Different researchers will have different opinions about which software is accessible, with some using only the most traditional point-and-click software, and others comfortable in more challenging environments. The SPSS macro (Marso and Shadish 2014) for the Hedges et al. (2012, 2013) effect size is quite accessible to SCD researchers given that SPSS (IBM Corp. 2013) is probably the software of choice for most of them. It is also the only software to incorporate power analyses for SCD designs. Some SCD researchers also work in the R programming environment (R Development Core Team 2012), which has the great advantage of being free. For them, the Pustejovsky et al. (2014) R package will be attractive. The BUGS environment used by the Swaminathan et al. (2014) approach, which is also free, may be least accessible to most SCD researchers, but accessibility may improve as those authors finish development of their own specialty software package for SCD researchers.

To summarize these comments, here are the major characteristics to consider for each between-case effect size:

- The Hedges et al. (2012, 2013) effect size
    - The simplest to compute, uses SPSS macro with graphical user interface.
    - The researcher should test for normality of residuals. If the test does not support normality, report this in results.
    - Test for presence of trend statistically or visually (or both). If trend is likely, use the detrend option in the macro to see if it yields the same effect size as without that option. If they are similar, trend may be less a concern. If they are different, report and discuss.

- The Pustejovsky et al. (2014) effect size
    - Uses the R program environment.
    - Allows testing random effects.
    - Also assumes normally distributed residuals.
    - Use if trend is clearly present and would change the effect size.
- The Swaminathan et al. (2014) effect size.
    - The most flexible of all in distribution theory, trend, and random effects.
    - Software not widely available yet; model syntax in the BUGS language is available in Swaminathan et al. (2014).

3.4 Making No Choice Is a Bad Choice: The Perfect as the Enemy of the Good

The fact that none of these between-case effect size estimators meets all evaluative criteria may discourage the reader from choosing and using one. That would be unfortunate for two reasons. The main reason is that many of the flaws in these effect sizes are relatively minor and will no doubt be remedied in the not-too-distant future. The priority should be on using them to help encourage the inclusion of SCDs in evidence-based practice reviews. Continued failure to report effect size statistics will only continue the practice of excluding SCDs in systematic reviews. The second reason is that no alternative effect size indictors are better. For instance, all the within-case effect size estimators generally do not take autocorrelation into account, have little firm grounding in statistical theory, mostly ignore trend, and lack developed power analyses (Shadish, 2014b). Also, with the exception of some of the overlap statistics,[23] few of the other within-case effect size estimators have user-friendly interfaces, although those who find the R computing environment to be accessible will find the Single-Case Data Analysis (SCDA) package very useful[24] (Manolov, Gast, Perdices and Evans 2014)

In the end, the SCD researcher should choose one of these effect size estimators based on its match to the needs and characteristics of the data. It is always possible to compute more than one effect

[23] http://www.singlecaseresearch.org/calculators.

[24] http://www.researchgate.net/profile/Rumen_Manolov/publication/275517964_Single-case_data_analysis_Software_resources_for_applied_researchers/links/553e04c30cf29b5ee4bcf6fb.pdf.

size (or indeed, more than one analysis!) on the same data, of course, but we would greatly increase the chances of including SCDs in evidence-based practice reviews if we simply chose one of these between-case effect sizes and reported it.

4. How to Report Between-Case Effect Sizes in Single-Case Designs

Reporting effect sizes is not intended to replace visual analysis in SCD research because visual analysis is a process used throughout an SCD study from design to publication and because SCD researchers often report at least some statistical analyses (e.g., reporting means or percentages by phase) in addition to their visual analysis. Rather, reporting effect sizes is intended to complement visual analysis by providing evidence-based practice reviewers with a familiar touchstone through which to assess the contribution of the study.

To facilitate the use of SCD research in evidence-based practice reviews, whether new reviews or revisions of previous reviews, we suggest the following four reporting practices:

1. Report a between-case standardized effect size, standard error, and inferential test.
2. Report citations to the between-case method used and to associated software or syntax.
3. Report assumptions of the effect size used and results of any tests of those assumptions.
4. Make raw numerical outcome data available.

Next we show specific examples of how to implement these four practices.

4.1 Report a Between-Case Effect Size, Standard Error, and Inferential Test

4.1.1 Example. Tasky, Rudrud, Schulze, and Rapp (2008) used an ABAB design to study whether on-task behavior would increase if they allowed three women with traumatic brain injury to choose which tasks to do. In the baseline (A) condition, the women were assigned three tasks at random. In the treatment (B) condition, they were allowed to choose three tasks from a list of nine tasks. The researchers measured the outcome, on-task behavior, using a whole-interval recording procedure. This involved observing the behavior of each case during the last 10 seconds of each 5-minute interval during a 30-minute period, scoring each interval as an occurrence only if the woman was engaged in on-task behavior the entire 10 seconds. The number of time points observed for the three women was n = 21, 24, and 25, respectively, for a total of N = 70 data points.

Shadish, Hedges and Pustejovsky (2014) conducted a secondary analysis of the Tasky et al. (2008) study, computing their between-case effect size and associated statistics on digitized data from this study, and reported that

> g = 1.605 and $s_g^2 = .103$ ($s_g = .320$). One can test the statistical significance of this effect size either with a z statistic ($z = 1.605/.320 = 5.02, p < .001$), or by seeing if the 95% confidence interval ($0.977 \leq \delta \leq 2.233$) excludes zero. By both tests, the effect is significant (p. 534).

Of course, this simple quotation is considerably elaborated in Shadish et al. (2014), as it would be in any such publication, but the quotation suggests some minimum prose and statistics as examples on which to build those elaborations.

The rationale for this recommendation is that it will facilitate the inclusion of SCD research in evidence-based practice reviews for studies to report a standardized between-case effect size, its standard error (or variance, from which the standard error can be computed as the square root), and either a confidence interval or a significance test. The confidence interval is preferable, and if it is a 95% confidence interval, then when it excludes zero from the interval, the inferential conclusion about the significance of an effect should be the same as from a significance test at the $\alpha = .05$ significance level. Using the between-case standardized effect size allows results from SCDs to be compared to results from between-groups designs that are so common in most evidence-based practice reviews. Those reviewers may or may not choose to combine studies using different designs, but between-case effect sizes allow them to do so legitimately if they so choose.

Such reporting would be consistent with current reporting practices. For example, the WWC is the main institution for identifying evidence-based practice in education. Its *Procedures and Standards Handbook* (U.S. Department of Education 2014a) often discusses the usefulness of significance testing, effect sizes, and confidence intervals. The handbook, for example, tells reviewers to take "into consideration the number of studies, the sample sizes, and the magnitude and statistical significance of the estimates of effectiveness" (p. 2) and that "to adequately assess the effects of an intervention, it is important to know the statistical significance of the estimates of the effects in addition to the mean difference, effect size, or improvement index" (p. 24). In addition, significance

testing was recommended by the APA Task Force on Statistical Inference (Wilkinson, and Task Force on Statistical Inference 1999), which encouraged researchers to report effect sizes, confidence intervals, and statistical significance tests. Significance tests should not be used in isolation to recommend evidence-based practice, but wide support exists for them as one part of a larger process.

Some SCD researchers may do other statistical analyses (or not) on their data. Doing so can have many advantages in addressing some of the limitations of between-case effect sizes. For example, multilevel models can quantify the extent to which treatment effects vary over cases within a study (Shadish, Kyse and Rindskopf 2013); generalized additive models can model nonlinear trends better than most alternatives (Shadish, Zuur, and Sullivan 2014); and many different kinds of regression analyses can model the effects of more than two conditions at a time (e.g., simultaneously modeling outcomes from baseline and two alternating treatments A and B). However, such statistics cannot substitute for reporting a between-case effect size because it is rarely clear how such an effect size could be computed from reports of other analyses. For example, no valid method currently exists for converting a set of overlap statistics or the results of a within-case regression analysis to a between-case effect size, at least not without access to the raw data and specialized analytic methods.

4.2 Report Citations to the Between-Case Method Used and to Associated Software or Syntax

4.2.1 Examples. Smith, Eichler, Norman, and Smith (2014) studied the effectiveness of collaborative/therapeutic assessment for psychotherapy consultation using SCD. They analyzed their data several ways, one of which was to compute a between-case effect size. They reported that

> "We calculated the *d* statistic to garner the overall magnitude of effect for the intervention (Hedges, Pustejovsky, and Shadish 2012; Shadish, Hedges, Pustejovsky, Rindskopf et al. 2014). We used the DHPS macro (Shadish, Pustejovsky, and Hedges 2013) in SPSS Statistics (2012) for this analysis." (p. 6)

Swaminathan et al. (2014) reanalyzed data from an SCD study reported by Lambert, Cartledge, Heward, and Lo (2006) on the effects of response cards on disruptive behavior and on responding to questions a teacher asked during math lessons by fourth-grade urban students. They reported that

"The analyses were carried out in OpenBUGS (Lunn et al. 2009). The input file for the analysis along with the data and other necessary specifications are provided in the Appendix" (p. 222).

The three between-study effect sizes differ from each other in details such as whether and how they deal with trend, the assumptions they make about the outcome variable distribution, and how they deal with serial dependence among outcome observations within cases. Citation to the specific method and the software or syntax used to compute the effect size allows the reader to obtain the sources needed to identify those details. Sometimes providing that information can be done by citing published or otherwise available primary sources, as when an author cites previously published syntax used for a similar analysis. Alternatively, researchers can accomplish the same goal by publishing the syntax used or making that syntax available as supplementary material.

4.3 Report Assumptions of the Effect Size Used and Results of Any Tests of Those Assumptions

4.3.1 Example. Shadish, Hedges, and Pustejovsky (2014) analyzed the Lambert et al. (2006) SCD study with a between-case effect size that assumed normally distributed residuals about the phase means within persons and no trend. For trend, they reported that

"We can test the assumptions of stationarity and normality. When we tested the Lambert et al. (2006) data for trend using multilevel modeling in a previous publication (Shadish, Kyse, and Rindskopf 2013), trend was nonsignificant. Both Moeyaert (2014-this issue) and Shadish, Zuur, and Sullivan (2014-this issue) found more ambiguous results about trend, in particular, the possibility that trend might be present in some phases but not others or that trend might be nonlinear. Given that, we also computed *d* using the detrending option in the macro." (pp. 129-132)

For normality, they reported that

"For the Lambert et al. (2006) data, the Shapiro-Wilk test rejected normality, the normal quantile-quantile plot was ambiguous, and the detrended normal quantile-quantile plot suggested the data were not normal." (p. 131)

The rationale for this recommendation is that potential readers and users of SCD research will want to know what assumptions are required in order to validly use the effect size estimator, so they can decide for themselves whether those assumptions are realistic and whether potential violations of them are likely to be consequential. To be even more helpful, SCD researchers can statistically test and report the validity of those assumptions so the reader does not need to rely on visual inspection of SCD graphs to make that judgment.

Users of these statistical tools may worry that transparent reporting of assumptions and their tests will discourage the use of SCD studies when assumptions might be seen by others as implausible or violated. We believe this concern is overstated for four reasons. First, these assumptions can be feasibly tested, and sometimes dealt with, as when data are detrended before analysis. Section 3.3 showed how to do so in many instances. Second, honest reporting of this sort increases the credibility of the report. Third, such reporting might be more than is often reported in between-case studies that, for example, also assume normality but do not test the assumption. Fourth, evidence about the consequences of violation of assumptions is meager at this point, with some evidence suggesting the possibility of robustness to violations (e.g., Shadish 2014a; Shadish, Hedges and Pustejovsky 2014). This is not to trivialize such violations, but transparent reporting of them may be one of the prompts needed for researchers to develop effect sizes with fewer or different assumptions or evidence about the consequences of their violations.

4.4 Make Raw Numerical Outcome Data Available

4.4.1 Examples. Schutte, Malouff, and Brown (2008) used a multiple-baseline design to study the efficacy of an emotion-focused therapy in 13 adults suffering from prolonged fatigue. That study illustrates one way in which SCD researchers can make numerical outcome data available, namely, by publishing the raw data for each person and each session in a table in the publication, as when Schutte et al. (2008) said:

> "Table 1 shows baseline and treatment fatigue severity scores for each participant" (p. 706).

Shadish, Hedges, Pustejovsky, Boyajian, et al. (2014) reanalyzed data from the Tasky et al. (2008) study of an intervention with women suffering from traumatic brain injury. That study illustrates a second way in which the SCD researcher can make numerical outcome data available, namely, by

publishing the raw data in an appendix to the publication, as when Shadish, Hedges, Pustejovsky, Boyajian, et al. (2014) said:

> "The data are in the appendix." (p. 534)

A third way to accomplish this recommendation would be to make the data accessible as supplementary material in an online format, but we have not found a SCD study that actually did so that we could cite as an example here.

SCD results are nearly always reported in graphical form, with time on the vertical axis, outcome on the horizontal axis, and dots representing outcomes in the body of the graph where the dots are connected by lines. That practice of sharing raw data is commendable and should continue for all the reasons SCD researchers have always cited. However, researchers who wish to reanalyze such data must first digitize it. Digitization can be highly reliable and valid (Shadish et al. 2009), and free programs for doing so exist, such as the PlotDigitizer (http://plotdigitizer.sourceforge.net/) or the SCDA plug-in for R (http://cran.r-project.org/web/packages/RcmdrPlugin.SCDA/index.html) (Bulté and Onghena 2012). However, the Shadish et al. (2009) study also reported that the persons doing the digitizing require extensive training and ongoing monitoring, the process can be time consuming, and some graphs are not drawn in a manner that facilitates easy digitization (e.g., observations are too densely packed to be separately identifiable; multiple outcomes are reported in one graph but are not clearly distinguishable). It would be far easier, and even more accurate, if the raw data themselves were available in tables, appendices, or supplementary materials online, so the analyst did not have to digitize them.[25] Archiving data as part of supplementary materials is increasingly the norm for much of science and is entirely feasible for most of the publication outlets used by SCD researchers. So this should become (and is already becoming) part of good scientific practice in all fields, including SCD research.

Raw data should be available even if the researcher reports an effect size and all the other material outlined above. The reasons are that (a) scientific ethics often call for the release of raw data for purposes of verifying published results (e.g., see American Psychological Association, 2002, standard 8.14), and (b) in the future, we cannot know what effect sizes (or other analyses) are likely to

[25] A fourth option is for the reanalyst to contact the original author and request the data, but this can be problematic for older studies. Shadish and Sullivan (2011) reported no success in obtaining raw data from a small set of past SCD researchers they were able to locate.

become the eventual consensus standard as the limitations of current effect sizes are addressed, and so we must anticipate the needs of future evidence-based practice reviewers to compute their own effect sizes (or other analyses) on the raw data.

The dataset should include certain minimum information: (1) the numerical outcome for each case at each time point, reported at several decimals of precision if the outcome is continuous;[26] (2) the session number and/or chronological time[27] associated with each outcome observation; (3) an identifier for each case; and (4) an identifier for each outcome measure when more than one outcome measure is reported. Exactly how this is done will vary depending on the format used to report the data. For example, if reported in a publication table (Schutte et al. 2008, is a reasonable example to imitate), the case identifier might be the (typically fictional) name of the case, and the outcome identifier might be a narrative label consistent with the words used in the *Methods* section of the article. Or, if the data are reported in an appendix or supplementary material (Shadish, Hedges, Pustejovsky, Boyajian, et al. 2014, is a reasonable example to imitate), the data set might contain one line of data for each time point for each case, with each line listing a numerical[28] case identifier, a numerical outcome identifier, the session number or chronological time value, an indicator for which treatment condition (e.g., baseline = 0, treatment = 1) applies to the observation, and the numerical outcome.[29] When reported in this latter, entirely numerical fashion, the original researcher can also provide a brief narrative description linking the numbers to the words used to identify cases and outcomes in the study report.

As it becomes normative for SCD researchers to save and archive their data for requests from others, additional reporting norms may prove desirable. An example would be reporting potential

---

[26] How many decimals to report will vary depending on the outcome measure. For example, weight may often be reported to one decimal (a half pound or kilogram) due to measurement devices, but reaction time measured by a computer might be reported to many decimals if those times are less than one second in a substantial portion of sessions.

[27] For statistical reasons, chronological time (e.g., days or fractions thereof) is preferable to session number, but the details about why are beyond the scope of this paper. Of course, both session number and chronological time can be reported in the dataset even if the graphical presentation uses session number for convenience.

[28] Numerical identifiers save the analyst time having to convert narrative identifiers to the numerical ones most analysis programs will require.

[29] When the outcome is reported as a rate or proportion (e.g., the proportion of trials in which the outcome was observed within a session), the dataset should have another column that indicates the number of trials per session (separately for each session if not constant over sessions).

covariates that can be used as moderating variables (e.g., age of child, diagnostic test results) that might help to understand variability in effect sizes. This would allow the researcher to compute a between-case effect size separately for cases in each category of the moderator (e.g., separately for males and females) so long as each category has sufficient cases (e.g., three in each category for Hedges et al. 2012, 2013, effect sizes). Even when a study has no variation in a moderator (e.g., all cases are males), reporting such values allows examining the effect of the moderator meta-analytically across studies (e.g., comparing results for studies using males to results from similar studies using females). In the present article, we do not list this as a necessary requirement but only as one that will benefit from further thought from SCD researchers.

5. How to Use Between-Case Effect Sizes With Individual Single-Case Design Studies

The primary reason to use between-case standardized effect sizes in SCD analysis is that they generate effect sizes consistent with those used in analysis of between-groups studies, allowing valid comparison of effects from SCD to group experiments. However, using those effect sizes will have three other benefits as well: (a) improving design sensitivity, (b) comparing results from visual and statistical analysis, and (c) accumulating descriptive data about the magnitudes of effect sizes in SCD research.

5.1 Improving Design Sensitivity

Design sensitivity (Lipsey and Hurley 2009) consists of (1) power, or the probability of correctly concluding an effect is present when there is an effect; (2) precision, or the width of the confidence intervals around the point estimate of a treatment effect; and (3) minimum detectable effect size, or planning a study that is capable of detecting the minimum effect size judged to be important before a study is implemented.[30] Effect sizes allow SCD researchers to assess all of these aspects of design sensitivity, which are inseparably intertwined.

The value of attending to design sensitivity stems directly from the same rationale that justifies reporting between-case effect sizes in SCDs, that is, the connection to evidence-based practice reviews (whether meta-analytic or not). Presumably, SCD researchers want their research to be included because that research provides valid evidence about treatment effects. If so, SCD studies that are underpowered[31] are likely to have very wide confidence intervals and to show that the effect size is not statistically significant. Ironically, then, SCDs would be included in evidence-based practice reviews but would be used to show a lack of evidence for a treatment effect—hardly the outcome that SCD researchers have in mind. This problem can be partially remedied using meta-

[30] Minimum detectable effect size is closely related to power, but power is expressed as a probability, such as .80, not as an effect size. Further, the power of a study may allow detecting an effect size that is considerably larger (or smaller) than the minimum detectable effect size.

[31] If the Type II error rate is $\beta = .20$, it is common to use $1 - \beta = .80$ as a power cutoff, implying 80 percent probability of finding a significant effect if there is an effect in the population. As in any literature, this is a cutoff established by professional consensus, grant proposal expectations, and the needs of the context, and so it can be changed given a compelling rationale.

analysis to improve power by pooling results from many underpowered studies, and we discuss this option in the next chapter. Power analyses for such meta-analysis are already well developed (Hedges and Pigott 2001, 2004). Even so, designing sufficiently sensitive studies in the first place, when it is feasible, is a better option.

In between-groups studies, power is substantially (but not exclusively) affected by sample size, that is, by the number of cases in the study. By analogy, SCD researchers may worry that SCDs tend to have such small numbers of cases that power will always be a problem. However, focusing only on the number of cases can be misleading, especially for SCDs. Power is really affected by the amount of information in a study. More cases provide more information, but so do more observations per case, more phases and phase reversals, and more observations with smaller serial correlations over time. The total information in many SCD studies is often quite a bit higher than the number of cases by itself might suggest.

Still, some researchers may worry that it is not feasible to design sufficiently powerful SCD studies and especially that increasing power by increasing the number of cases or observations may not be feasible. It is true that many SCD researchers operate under logistical and practical constraints that prevent them from gathering more data (c.f., Barlow et al. 2009). Further, educational and clinical needs play a role in the design of an SCD study. In some clinical settings, the problem is so severe that the intervention should start quickly, requiring a shorter baseline than desirable. In such cases, power may be lower than theoretically desirable. Power analysis helps those researchers know what the power of a study might be and qualify their statistical conclusions as needed.

Yet such situations are probably the exception rather than the rule (Shadish and Sullivan 2011). The majority of SCDs in psychology and education (Shadish and Sullivan 2011; Smith 2012) study problems where acute risk is not at issue, as with many positive behavior interventions, or problems like communication deficits related to ASD where the SCD researcher often has some leeway to gather more data. Many SCD researchers may simply be unaware that the number of cases or observations in a planned design may yield an underpowered study, that is, a study that is unlikely to detect an effect as statistically significant when an effect is present. Some recent methodology standards for SCDs (see Smith 2012, for a review) encourage minimum numbers of observations that exceed what has been common in much SCD research (Shadish and Sullivan 2011), and those

standards often encourage more observations than the bare minimum that would otherwise meet standards (e.g., Kratochwill et al. 2010). For instance, the WWC Pilot Standards suggest having a minimum of 20 observations in an ABAB design, 5 in each phase. That number itself is actually only the median in SCD research, with half of the studies having fewer than that, but the WWC standards also encourage using even more than that minimum of 20 where feasible.

Power analyses can quantify exactly how many more observations might be desirable in order to detect a minimum effect based on substantive knowledge of practically important effects or on relevant past research. Here are some examples from preliminary work by Hedges and Shadish (2015a, b), though these results can be replicated using the power macros accessed in the box for worked-through examples for Hedges et al. (2012, 2013) in section 3.3. All examples here assume an autocorrelation of 0.50 and a ratio of between-case variance to within-case variance of 0.50, reasonable starting points for now.

1. For ABAB designs, which tend to have more power than multiple baseline designs:
    a. With five cases and three observations per phase, power will exceed .80 if the between-case effect size is $\delta > .70$.
    b. If the effect size is only $\delta = .50$, then nine cases is always sufficient for power to exceed .80, and six cases will do so with five observations per phase.
2. For multiple baseline designs:
    a. If at $\delta = 1.00$, then four cases will yield power of at least .80 if there are 8 observations per phase, but with 3 observations per phase, even 12 cases will not yield adequate power.
    b. If at $\delta = 1.00$, with five observations per phase, at least seven cases are needed to obtain power of .80.

Clearly the expected effect size (δ) makes an enormous difference to power so that gathering more information about common between-case effect size levels in different areas of SCDs is crucial (see section 5.3). The effect sizes in table 3 and the SCD Spelling Mastery study effect size in section 6.1 suggest that effect sizes about $\delta \approx 1.00$ may not be an unrealistic expectation, though clearly more data are needed. Of course, these power estimates apply only to the Hedges et al. (2012, 2013) effect

size, and when other effect sizes try to incorporate trend or random effects into the computations, power will decrease, and more cases or observations will be needed.

When it truly is not feasible to gather more observations or cases, SCD researchers can use other methods to improve power (Lipsey and Hurley 2009) such as improving reliability of the outcome, increasing treatment dose, increasing the number of trials per session, timing the measurement of effect size to correspond to maximum treatment effect, and using different Type I or Type II error rates than the usual conventions. The latter follows the lead of Baer (1977), who suggested that SCD researchers might be willing to risk overlooking some effective treatments (that is, increasing the Type II error rate above the usual $\beta = .20$) in the service of decreasing the number of false positives, suggesting a treatment works when it does not (that is, lowering Type I error to something more stringent than the usual $\alpha = .05$, say, $\alpha = .01$). However, just having this discussion requires a measure of effect size for SCDs with known distribution properties and standard errors, which most of the between-cases effect sizes have.

Even when none of the above approaches are sufficient to increase power of an individual SCD study to .80, results from that study may still be input into a meta-analysis. Power for a meta-analysis of many studies, each of which may itself be underpowered, may be quite high. Section 6.2 discusses the use of meta-analysis in detail. The point of power analysis is not to show that an intervention is effective, but rather to ensure that the design allows a reasonable probability that an effect of a certain size could be detected if it is present. A statistically significant effect may not be present, and that knowledge is just as important to evidence-based practice reviews as is finding a significant effect.

5.2 Comparing Results From Visual and Statistical Analysis

In SCD studies, effect sizes and visual analysis can be used jointly to better understand the effects of an intervention. For instance, professional wisdom among many SCD researchers (e.g., Baer 1977) is that visual analysis of SCDs yields fewer Type I errors but more Type II errors—that when they say there is an effect, we can surely believe it. Effect sizes will allow SCD researchers to explore such assertions. One could envision a study that takes a sample of published SCDs to record the conclusions reached by the authors about whether the treatment was effective, and then computes

an effect size and significance level on those same studies. This process would allow comparing visual to statistical results to shed light on the error rate problem. This need not require SCD researchers to give priority to visual over statistical analysis or vice versa. Rather, it is a way to provide a new perspective and to open a dialog about assumptions that are important to the field.

Of course, computing an effect size, doing a visual analysis, and comparing them to each other, is not sufficient to make a statement about Type I or Type II errors. The underlying characteristics of the data (i.e., whether a true effect exists in the population) need to be known (as in a simulation study). Neither the value of the effect size nor the inferences drawn in a visual analysis can be assumed to represent this truth perfectly. Such studies will always provide clues rather than definitive answers.

To emphasize again, the purpose of comparing visual to statistical analysis is not that effect sizes should replace visual analysis. Our expectation is that the SCD researcher will continue to use visual analysis throughout the conduct of the SCD study. Visual analysis in the SCD tradition is not just a tool for making a summative judgment of the outcome of treatment for a case (Perone 1999). Rather, it is used formatively throughout the conduct of the study to examine such matters as the nature of the pattern of baseline responding, when to initiate or terminate a phase, whether the treatment of a case needs adjustment to improve outcome, and whether a case seems to be responding at all. The logic of causal relationships in SCDs relies on visual analysis of how the case responds to the various experimental manipulations, not on statistical analysis. The standardized effect size is simply a useful statistical summary of what the SCD researcher sees in the summative visual analysis of the final graphs, much in the same way that a correlation coefficient is a useful statistical summary of a scatterplot.

5.3 Accumulating Descriptive Data About Effect Sizes

Another benefit of effect sizes to both statistical and applied researchers would be the accumulation of a descriptive database of effect sizes and associated statistics (autocorrelations) across SCD specialties, and comparing SCDs to other designs like between-groups experiments. The classic small, medium, and large categories were popularized by Cohen (1988), but the cut points he set (d = .2, .5, and .8, respectively) should not be accepted uncritically for use in SCD research (even

Cohen would not have recommended such uncritical use). One reason is that effect sizes in SCDs may vary by field within the larger SCD community or may vary by many other factors, such as the length of the intervention, just as they vary by such moderators in between-groups research. For example, SCD interventions that use applied behavior analysis to influence outcomes for children with ASDs (Rao, Beidel, and Murray 2008; Shadish, Hedges, and Pustejovsky 2014) may yield systematically different effect sizes than interventions to affect outcomes in neuropsychological rehabilitation (Shadish, Hedges, Pustejovsky, Rindskopf et al. 2014), cognitive rehabilitation (Cicerone et al. 2000, 2005, 2011), attention deficit hyperactivity disorder (Fabiano et al. 2009), or writing interventions (Rogers and Graham 2008). The systematic reporting of effect sizes (and of the raw data that allow computation of effect sizes) will kick-start the long-term task of accumulating empirical data about magnitudes of effects across SCD research and of pinpointing more accurately the degree of effect size that may affect social validity measures (see footnote 4 about the nature of social validity measures). This is not to imply that effect size is positively correlated with social validity, of course. In many areas, even very small effects from a quantitative point of view may have great importance socially, and that is true in both SCD and between-groups research.

For example, Shadish and colleagues have used the Hedges et al. (2012, 2013) *d*-statistic to meta-analyze SCD studies on three different topics: neuropsychological rehabilitation (Shadish, Hedges, Pustejovsky, Boyajian et al. 2014), ASD (Shadish, Hedges, and Pustejovsky 2014), and the effects of choice-making on challenging behaviors in people with disabilities (Zelinsky and Shadish 2014). Table 2 summarizes the results.

Table 2. Between-case effect sizes from meta-analyses on three different single-case design topics

Topic	Bias corrected effect size	Standard error
Pivotal Response Training in Autism	1.01	0.14
Choice Making With Disabilities	1.02	0.17
Neuropsychological Rehabilitation	1.27	0.24

Here, the effect sizes are fairly similar, as are the standard errors, but this is only on three of the many topics SCD researchers study. A much larger sample would allow SCD researchers to construct empirically based norms for the size of effects in different areas. Both statistical and

applied researchers would benefit from such descriptive databases. Statistical researchers would benefit from this kind of information by better understanding the likely statistical power of SCDs to detect effects. If effect sizes like those in table 2 prove commonplace, the power computations in section 5.1 of this paper might suggest the need for fewer cases (or fewer observations within cases) than was the case in those analyses. Applied researchers would benefit as well. For example, an applied researcher could compare an effect size for a new intervention to field-specific norms and, thus, know if the new intervention produced small, medium, or large effects compared to the effects of past treatments, and the researcher could better plan sample sizes, numbers of observations, and other decisions that hinge at least partially on effect size.

6. How to Use Between-Case Effect Sizes to Identify Evidence-Based Practices With Single-Case Designs

This chapter will discuss and demonstrate some benefits of using between-case effect sizes for SCDs in the context of evidence-based practice reviews. First, we discuss what would have happened if a recent WWC report had included an SCD study that it had excluded (for good reasons). Second, it discusses how researchers can incorporate modern meta-analytic statistics into reviews that use SCDs in ways that may improve the reviews. Third, it discusses conceptual issues that arise in considering whether results from SCDs should be combined with results from between-groups studies in evidence-based practice reviews.

6.1 An Example of Including SCD Results in an Evidence-Based Practice Review

The WWC recently published an intervention report on the effects of the Spelling Mastery curriculum for students with learning disabilities (U.S. Department of Education 2014b). Appendices C and D of that report present its effectiveness data, including between-groups standardized mean difference statistics (and associated significance tests) based on data from the two randomized experiments testing the curriculum, noting that "the WWC uses a standardized measure to facilitate comparisons across studies and outcomes" (p. 15). One of the two experiments, for example, used an author-created spelling test and two different subtests from a previously published spelling test. The other experiment used the same published spelling test but also three different author-created tests of the generalization of learning gains to other words, other contexts, and other times. For illustration, table 3 shows the effect sizes from appendix C of the WWC report.

Table 3. Effects of the Spelling Mastery curriculum

Experiment	Effect size	*p*-value
Experiment 1		
Spelling test (author-created)	1.42	< 0.01
Published spelling subtest 1	1.20	< 0.01
Published spelling subtest 2	1.18	< 0.01
Experiment 2		
Published spelling test	0.39	0.14
Generalization test (author-created)	0.43	0.10
Transfer test (author-created)	0.41	0.12
Maintenance test (author-created)	0.46	0.08

In the first experiment, the effect sizes are large, over one standard deviation improvement, all statistically significant. In the second study, the effects are somewhat smaller, less than half a standard deviation, all not significant. The average (unweighted) effect size over outcomes and experiments was a nonsignificant $d = 0.85$. Without a standardized effect size, it would not have been possible to compute that average effect over such diverse outcome measures.

The WWC considered whether to include an SCD study that had examined the Spelling Mastery curriculum (Owens, Fredrick, and Shippen 2004). The report did not include that study in the final analyses because it did not meet the pilot WWC single-case design standards due to having fewer than three data points in a phase. While the WWC did not include this study, it is possible to imagine that another review group might have done so. If they had, the issue that motivates the present paper is salient—without an effect size that is in the same metric as those in table 3, it would not be possible to compare the SCD results with those in table 3, nor combine them sensibly. Effect sizes that are standardized within case rather than between cases do not suffice for reasons outlined in section 3.1.2, including the overlap statistics. Only the three between-case effect sizes we outlined in section 3.3 can do this.

To illustrate, we computed the Hedges et al. (2012, 2013) *d*-statistic on the Owens et al. (2004) SCD data. That study reported results about the effects of Spelling Mastery on 6 cases with 2 outcomes

per case (percent correct letter sequences, percent correct words) using a multiple baseline across participants design over no more than 18 sessions total. For percent correct letter sequences, the small-sample corrected g = 1.165 (SE = 0.424), and for percent correct words, g = .890 (SE = 0.339), both statistically significant. Because the effect size is in the same metric as the randomized experiments in the WWC Spelling Mastery report, it can legitimately be compared to them or combined with them. The effect sizes from the Owens et al. (2004) SCD study are about the same magnitude as the first randomized experiment from the WWC report in table 3 above.

In fact, combining the Owens et al. (2004) effects with the effects from the two randomized experiments changed the overall nonsignificant average effect slightly to a significant g = 0.83 (SE = 0.26, p = .0012).[32] The point is not that the reviewers *should* combine results from SCDs with results from between-groups experiments. Rather, the point is that they *can* do so and that doing so can make an important difference to results of an evidence-based review decision that the program worked or not (in this case, results went from nonsignificant to significant results overall). We discuss issues in whether they *should* be combined in section 6.3.

This is an example of how to revisit previously completed reviews by adding SCDs using between-case effect sizes. Doing so requires either access to the original SCD data or digitization of data from graphs used in the studies in that prior review.

6.2 Using Modern Meta-Analytic Methods to Review SCDs

In the previous section, we used a WWC intervention report as an example. Yet formal organizations like the WWC neither are, nor want to be, the only ones doing evidence-based practice reviews. Another way such reviews are done is by researchers who conduct meta-analyses of study results on an intervention and then report them in journal articles, book chapters, or other appropriate professional outlets. Unfortunately, reviews of meta-analyses done on SCD research (e.g., Maggin, O'Keefe, and Johnson 2011; Shadish and Rindskopf 2007) indicate that those meta-analyses make little or no use of modern meta-analytic statistics. Meta-analysis is entirely consistent

[32] We used a random effects weighted average compared to the unweighted simple average used in the WWC report. The small sample bias corrected random effects weighted average for just the two studies in table 3 was g = 0.777 (SE = .405, p = .055).

with the emphasis placed on replication in the logic of SCD research—just extended one level higher to replication across studies.

In this section, we discuss what some of those statistics are and why they are useful. Topics include fixed and random effects statistical models, heterogeneity testing, specialty graphics such as forest plots and cumulative meta-analyses, diagnostic statistics and graphics, moderator analyses, and publication bias analyses. Although the paper will not be able to present these methods in detail, the interested reader can consult other publications that provide more detailed discussion, examples of their use, and syntax for doing them (e.g., Shadish, Hedges, and Pustejovsky 2014; Shadish, Hedges, Pustejovsky, Boyajian et al. 2014).

6.2.1 Issues in Computing the Average Effect Size

Glass (1976) proposed that computing and averaging study-level statistics such as effect sizes is what defined meta-analysis. The years following Glass's seminal work saw rapid development of better statistics for doing meta-analytic work (Hedges and Olkin 1985; Shadish and Lecy 2015). Four pertinent developments regarding the average effect size are to correct for small sample bias, to recognize that effect sizes within studies are not independent, to weight effect sizes according to the amount of information they contain (precision) when combining them, and to choose between a fixed and random effects model for computation.

Correcting for Small Sample Bias. The *d* statistic overestimates the population effect size when sample sizes are small. A correction for this bias is available for two of the between-case effect sizes in this report, and it never does harm to use it even when sample sizes are large. Especially given that SCDs often have small sample sizes, the correction has a useful impact on interpretation. Both the Hedges et al. (2012, 2013; Shadish et al. 2014) and Pustejovsky et al. (2014) effect sizes make this correction automatically in their software, and this could be done with the Swaminathan et al. (2014) effect size by adding syntax.[33]

[33] The issue is slightly more complex in this latter case because Swaminathan et al. (2014) used a Bayesian approach in which the application of this correction would require further study.

In the between-groups literature where this correction was first developed, *small sample* referred to the number of cases, for instance, the number of people assigned to a treatment or control group in a randomized experiment. Recall from section 5.1, however, that the key statistical issue is actually the amount of information a study provides. In between-groups research, group sample size is an important influence on information, but even there, other factors also affect information. An example is when nesting occurs, as when classrooms are randomly assigned to conditions but outcomes are measured at the student level. There, the amount of information in the experiment is reduced as the intraclass correlation increases (the ICC is a commonly used measure of the extent to which observations or their errors are dependent). Similarly in SCDs, the amount of information is not just a matter of the number of cases but also a function of the number of observations within cases, the number of phase reversals in ABAB designs (in SCD parlance, one version of the number of replications), the serial dependence (autocorrelation) of observations within cases, and the ratio of between-case variation to within-case variation. Statistical details aside, our experience is that SCDs often provide considerably more information (and precision) than one might guess given the number of cases. A corollary is that the correction for small sample size bias—while it should probably always be applied in SCD research—may make less difference than one might think based solely on the number of cases in the study.

Dealing with Multiple Effect Sizes From One Study. Many SCDs will measure more than one outcome, leading to multiple effect sizes (one for each outcome). While these multiple effect sizes may be important to detailed interpretation of the study itself, including them in a meta-analysis presents a minor problem. Most statistical analyses make an independence assumption[34] (that errors from one observation have no systematic relationship with those from another observation), and meta-analysis is no exception. Different effect sizes computed from the same cases within a study—as might occur if more than one outcome measure is reported—cannot be presumed to meet that assumption. The problem can be addressed in several ways that include using only one effect size per study in an analysis, using the average effect size from a study as the unit of analysis, doing a multivariate analysis, computing robust standard errors, or using some coding of the type of measure as a moderator variable. Not using one of these solutions results in an average effect size that is unduly influenced by studies with large numbers of effect sizes and that produces inaccurate

[34] This is not specific to SCDs or to meta-analysis but is true no matter what the design.

standard errors, confidence intervals, and significance tests. Shadish, Hedges, and Pustejovsky (2014) provide a worked-through example of how to aggregate effect sizes within studies using the averaging method. Zelinsky and Shadish (in press) provide a worked through example of using robust standard errors to address the problem in SCDs.

Weighted Average Effect Sizes. Statistical theory says that studies that provide more information about population parameters (i.e., that have more precision) should receive more weight than studies that provide less information (have less precision). As we discussed previously, in the between-groups literature, precision is heavily a function of the number of units in a group (sample size)—the more people in each group, the more information. So it has become customary in the between-groups literature to give more weight to studies with larger sample sizes. While one could weight effect size (*d*) by sample size directly (e.g., $w = N$, which is what the Schmidt-Hunter (2015) approach to meta-analysis does), a different approach called an inverse variance weight turns out to yield an average that has the smallest standard error. To understand it, remember that any single effect size will be an imperfect measure of the population effect size because of sampling error—that one measured the effect on a particular sample of people who do not perfectly represent the full population of people. For the usual between-groups *d*-statistic, the sampling variance is

$$v = \frac{n_1 + n_1}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)} \tag{2}$$

where n_1 and n_2 are the sample sizes of two groups in an experiment. If you play with some numbers, you will see that the first part of the equation, sample size, usually dominates the result, and you will see that large sample sizes result in smaller sampling variance (v) or less uncertainty about the interval in which the population effect is likely to fall. That is as it should be and is one reason why large sample sizes are useful. Notice also that v and n are strongly negatively correlated—the bigger the n, the smaller the v. That is also as it should be—the larger the sample size, the less uncertainty we have about the population effect. So in between-groups research, instead of weighting by N, we could weight using $w = 1/v$ (the inverse variance weight).

All this is more complicated in SCDs because sampling error for a between-case SCD effect size is a function of many more factors—not just number of cases but also number of observations within case, balance in the number of observations over phases, number of phases, autocorrelations, and the ratio of between case variability to within case variability (Hedges et al. 2012, 2013). So it would

make little sense to use the $w = N$ approach to weighting because N by itself is not a very good index of the precision of SCD results. So the solution is to weight SCD effect sizes by precision, by the amount of information they provide about the population effect size. More cases provide more information but so do more observations within case, equal numbers of observations over phases, more phases, lower autocorrelations, and lower between-case variability. So SCDs need to take all these other factors into account (e.g., see Hedges et al. 2012, 2013, for the exact forms this equation takes for SCDs in that specific case). Shadish, Hedges and Pustejovsky (2014) provide a worked-through example of how to create weights and use them in meta-analysis.

Fixed- Versus Random-Effects Models. Some details about how to compose weights for meta-analysis will depend on whether the analyst uses a fixed- or a random-effects analysis model. The choice between the two depends on the inference the analyst wants to make. Under a fixed-effects model, the researcher is concerned only about generalizing to studies that differ from the current studies solely by sampling error—that is, by who happens to be in the study. All else about the studies remains the same. Under a random-effects model, the researcher wants a larger generalization ability, to studies that may differ not just in who is in the study but in other ways like using different measures, different treatment variations, or different settings. The vast majority of evidence-based practice reviewers are interested in the kinds of generalizations represented by a random-effects model, and so that model has come to be the standard in most meta-analytic work. Shadish, Hedges, and Pustejovsky (2014) provide a worked-through example of how to use both fixed- and random-effects models in meta-analysis of SCD studies.

6.2.2 Heterogeneity Testing

In the WWC review of the effects of Spelling Mastery, the seven effect sizes in table 3 were all different from each other, ranging from $d = 0.39$ to 1.42. These differences occur for two reasons: (1) chance (sometimes called sampling error or sampling variability) and (2) systematic differences. Differentiating these two sources is important because the researcher cannot be expected to predict chance variation but could predict systematic differences using moderator variables (e.g., general, age, diagnosis, length of treatment). Fortunately, meta-analysis allows us to make this differentiation using heterogeneity testing.

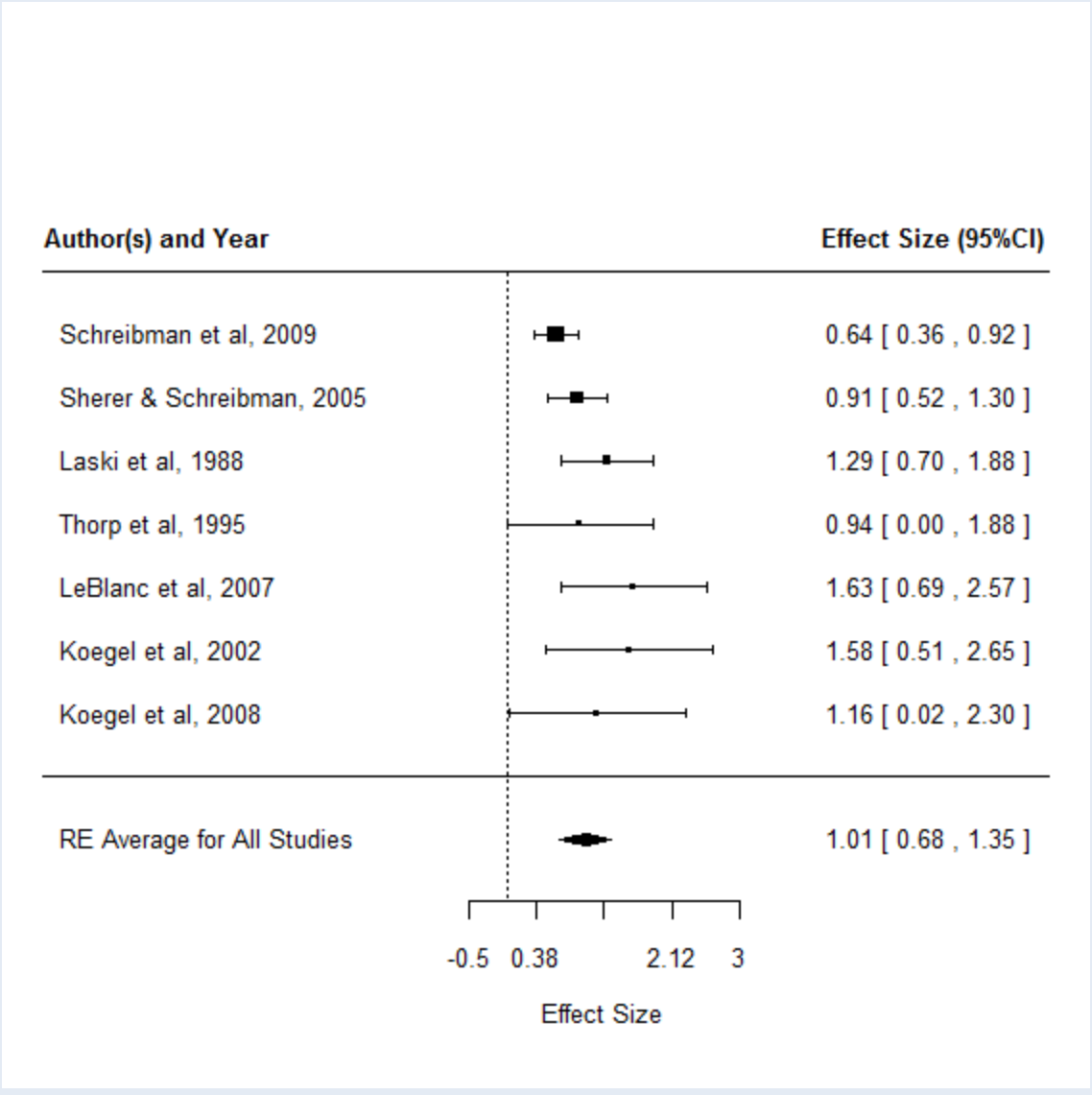
We can measure how much we expect a set of effect sizes to differ by chance using analytic formulas that measure sampling error associated with the effect size (Shadish and Haddock 2009). For example, the sampling variance of the usual between-groups *d*-statistic (equation (1)) is given in equation (2) above. To illustrate, if a randomized experiment produced a $d = .40$ and had sample sizes of $n_1 = n_2 = 30$ in each group, then $v = 0.068$. If we repeated the experiment but with different people, then we would not expect to see $d = .40$ but rather to see an effect size that varies from the original one due to sampling error. If we repeated the experiment a large number of times, we would expect to see a distribution of *d*'s with variance *v*. For example, the standard error of the effect size is the square root of *v*, or 0.261 in this example. We could compute a confidence interval around the effect size in the usual way, for example, resulting in a 95 percent confidence interval of $-0.11 \leq \delta \leq 0.91$. That is one way to indicate how much an effect size could vary by chance even when the underlying population $\delta = 0.40$—pretty big variation just by chance!

Now that we know how much variability we can expect by chance, we want to determine if the observed variance in a set of effect sizes is significantly larger than that and, if so, by how much. Meta-analysts use homogeneity statistics to test this. For instance, a *Q*-test statistic (essentially, a chi-square test) examines whether the amount of observed variance is significantly larger than would be expected by chance and an I^2 statistic that ranges from zero to one quantifies the percentage of observed variance in a set of effect sizes that is more than would have been expected by chance. In a nutshell, significant heterogeneity is a sign that the meta-analyst has more work to do in finding predictors of effect size variation. For example, for the three studies of Spelling Mastery (two in table 3 plus the SCD study), $Q = 2.95$ ($df = 2$, $p = .2285$), suggesting that the observed variability in the effect sizes over these three studies may not be greater than expected by chance. Like any other significance test, of course, this nonsignificant result might be due to low power (only three studies), a particular problem in meta-analytic heterogeneity testing where many meta-analyses have few studies (Shadish and Haddock 2009). The I^2 statistic is less influenced by sample size and in this case is $I^2 = .32$, so that about 32 percent of the total variation in these effect sizes is due to true variation in effect size parameters. This is usually considered a small I^2, but it does suggest some of the variation in effect sizes could be due to the systematic influences of moderator variables. Shadish, Hedges, and Pustejovsky (2014) provide a worked-through example of how to do heterogeneity testing in meta-analysis.

6.2.3 Forest Plots and Cumulative Meta-Analyses

Forest plots are graphs that display all effect sizes and their standard errors, and they are widely used in meta-analysis. Each row is a study effect size represented by a dark square, and a line depicting the size of the confidence interval surrounds the square. Figure 1 provides an example from a meta-analysis of seven SCD studies using Pivotal Response Training to help children on the autism spectrum (Shadish, Hedges, and Pustejovsky, 2014, provide a worked-through example of how to create these plots).

Figure 1. Sample forest plot



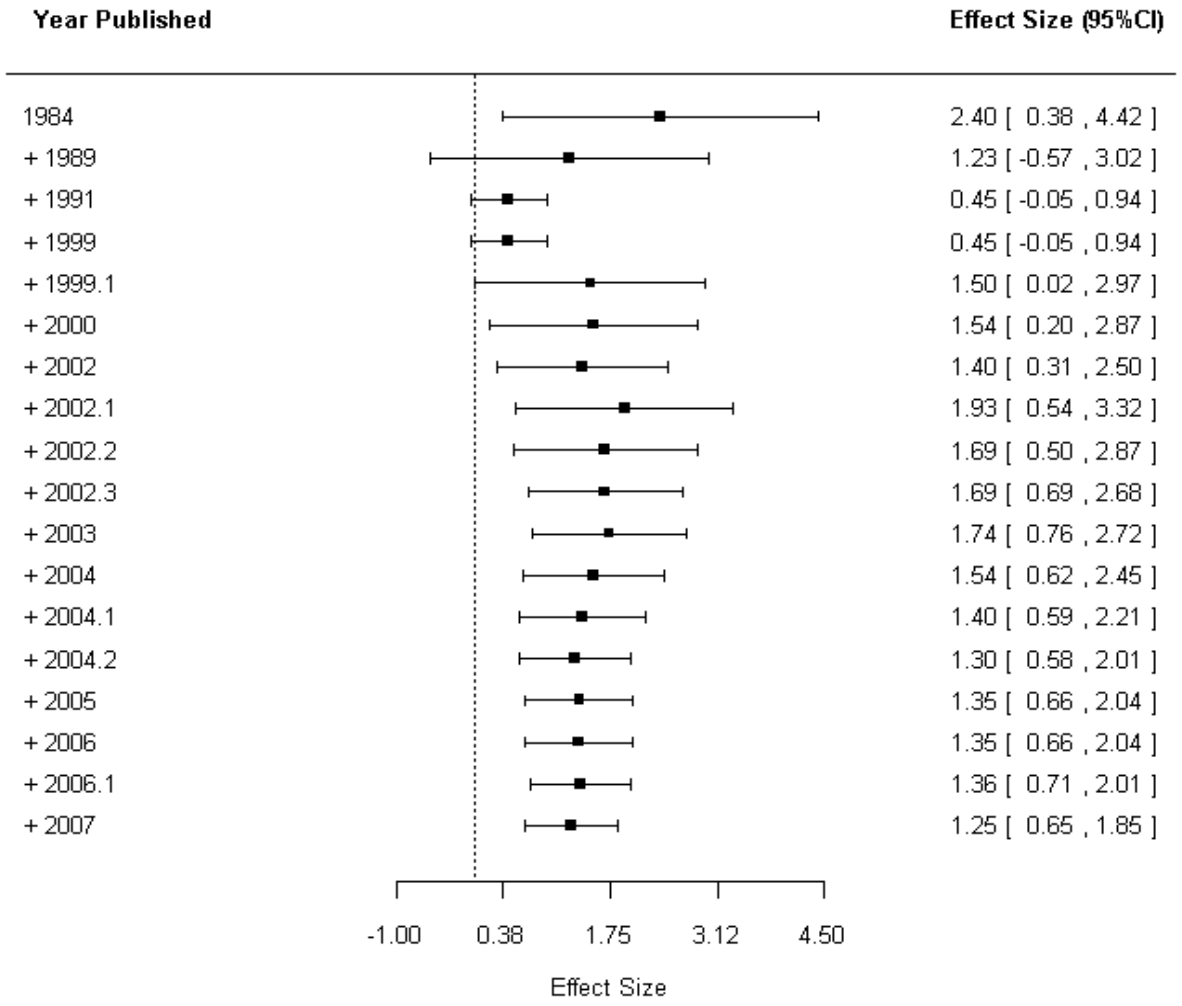
Many variations of forest plots exist. In figure 1, studies are ordered by study precision (smallest standard error); the size of the central square is proportional to precision; the line surrounding the square is a 95 percent confidence interval; and the random effects average is also presented. But studies could be ordered alphabetically or by year of publication; one could use wider or narrower confidence intervals; and one could either omit the meta-analytic average or present both a fixed- and random-effects average.

Individual SCDs sometimes lack sufficient precision to yield significant results, leading to large confidence intervals that indicate great uncertainty about the size of the population effect. That can happen even when the effect size is large, as with the Thorp et al. (1995) study in figure 1 with an effect size of 0.94 but with a confidence interval that includes zero. Pooling of results in a meta-analysis, as we have just described, greatly ameliorates the smaller power of single studies by taking advantage of the greater amount of information present across studies so that the confidence interval becomes quite a bit smaller for that meta-analytic average than for many of the individual studies themselves.

A cumulative meta-analysis (Lau, Antman, Jimenez-Silva, Kepelnick, Mosteller, and Chalmers 1992) is a forest plot that differs from the usual forest plot in two ways. First, it presents studies ordered by time from oldest to newest, and second, each row is the average of the study named in that row with all studies that preceded it in time (i.e., above it in the plot). The plot has several motivations. One is to make visually clear how the precision of an effect tends to increase over time as more and more studies are combined—usually, the confidence intervals around the average effect sizes tend to get smaller over time. Another is to suggest to reviewers when additional studies on the topic might or might not need to be done. Sometimes the cumulative average stabilizes to a value in a very narrow range with a very small confidence interval after a certain number of studies have been done so that new studies on exactly the same question may not be the wisest use of resources. A third motivation is to provide a method to policymakers who want to keep track of the cumulative results of the studies they have funded on a topic. For example, figure 2 presents a cumulative plot for a set of SCD studies of treatments in neuropsychological rehabilitation (Shadish, Hedges, Pustejovsky, Boyajian et al. 2014). The earliest study in that sample (done in 1984) had an effect size of $d = 2.400$. The average of the first and second studies ever done was $d = 1.23$, of the first three studies done

was $d = 0.45$, and so forth. The cumulative plot indicates how the average effect size changes as more studies are done. The effect size stabilized substantially as more studies were added, as evidenced by the location of the black squares stabilizing vertically in the graph. After about 2004, not much really changed in either the effect size or the confidence interval. Policymakers and researchers can both then ask if this program of research has matured in its current form, how future studies on the topic should be designed to yield novel information, or whether funding can be moved to other topics about which fewer studies have been done.

Figure 2. Sample cumulative meta-analysis



Given the purpose that instigated them, cumulative plots are always ordered from oldest to newest study. In theory, however, other orderings might make sense. For example, one could order studies from most to least precise, to see what happens to the meta-analytic average as less precise studies are added to the average. Such variations are not done mostly because meta-analysts have not found them very interesting, but they and other orderings are certainly possible.

6.2.4 Meta-Analytic Diagnostics

Diagnostic methods examine effect sizes for clues about potential problems. Complete sets of diagnostics are rarely reported in publications. Instead, the analyst uses them in an exploratory fashion to understand the characteristics of the data or whether the data meet the assumptions required for a meta-analysis. For example, leave-one-out methods recompute meta-analytic statistics (e.g., average, heterogeneity) leaving out each study one at a time to see how much the statistics are influenced by that study. Plots of standardized residuals can indicate if residuals are normally distributed. Galbraith plots suggest which studies contribute most to heterogeneity. When Shadish, Hedges, and Pustejovsky (2014, with worked-through examples) applied these methods to the PRT data in the forest plot presented earlier, one study (Schreibman, Stahmer, Barlett, and Dufek 2009) was consistently an outlier (with lower ES) and influential in the analysis. This does not mean something is wrong with this study—e.g., that study actually had the highest precision of all studies. Rather, the challenge for the meta-analyst is to decide why that case is so influential, a detective-like process likely to yield hypotheses for further research rather than certainty. For example, the sample of children with ASDs treated by Schreibman et al. (2009) was the youngest of all groups of participants in the PRT studies. It is possible that younger children obtain smaller benefits from PRT, but this remains a hypothesis for future work. An important contribution of measuring effect sizes may be identification of such possible moderating variables to be explored further in the meta-analysis (if enough studies varying on age exist) or in future research.

6.2.5 Moderator Analyses

Just as in any primary study, the meta-analyst can try to predict variation in effect sizes using the meta-analytic equivalents of *t*-tests, analysis of variance, or regression (Shadish, Hedges, and

Pustejovsky 2014, provide worked-through examples). The analyst codes a variable thought to influence effect size, for example, age of the children, and then uses that code to predict effect size. Just as is the case in computing the meta-analytic average, the analyses also involve inverse variance weighting so that studies with more precise estimates have more impact on the results than studies with less precise estimates.

Of course, one does not need to limit moderator analyses to the meta-analytic context. Individual study reports of SCD results can also use effect sizes to test moderators, for example, that the intervention works better for older children than younger children within the study. The main limitation of such an effort will be the lower power of that test in a single study given the few number of cases typically reported in any one SCD study. A meta-analysis is likely to have higher power for moderator tests.

6.2.6 Publication Bias Analyses

Publication bias refers to the possibility that the available studies (i.e., studies that the reviewer doing a meta-analysis has been able to obtain) are a biased sample of all studies conducted on a topic. Compelling evidence suggests that such biases do exist (Rothstein, Sutton, and Borenstein 2005). In between-groups research, the reason for this bias is that authors have a lower probability of submitting a manuscript for publication, and reviewers have a lower probability of recommending publication, if that manuscript reports nonsignificant results. If those studies are omitted from a review, the resulting estimate of the effect (whether that estimate is quantitative or narrative) may overestimate the effect that would have been estimated had all studies been available. Also, if those omitted studies are systematically different from the included studies in such ways as having different kinds of cases, settings, treatment variations, or outcomes, the field is deprived of knowledge about what variations do not result in the effect. After all, reliable knowledge about what does not work is also important for policymakers and practitioners to know.

At first glance, one might think that publication bias is unlikely in SCD research because so few SCD researchers use statistics, and without statistics one cannot reject a manuscript based on lack of statistical significance. However, publication bias might take a different form in SCD research. In many SCD communities, tradition requires demonstrating a large and visually detectable functional

relationship between the treatment and outcome (Kazdin 1982, 2011; Kratochwill et al. 2010, 2013), which seems to be the visual equivalent of a large effect size in statistics. The professional experience of the authors of this paper suggests that demonstrating a functional relationship is an important factor used in publication decisions about articles in the SCD community. If so, SCD researchers may (a) stop treating a case if treatment appears not to work and never present those results, (b) continue the study but decide not to write a manuscript or submit it for publication, or (c) include in a manuscript only cases that did find a functional relationship (although this would usually be considered inappropriate, some reason exists to think it sometimes happens[35]). Reviewers or editors may reject manuscripts without a demonstrated functional relationship. Reason exists to think at least some of these mechanisms operate in the SCD literature (Mahoney 1977; Sham and Smith 2014). So even if SCD researchers never calculate an effect size, their reliance using visual analysis to detect a strong functional relationship may lead them to give publication preference to studies that show larger effect sizes.

Regardless of the source, if there is a lower probability of finding studies with smaller (as opposed to larger) effect sizes in the available literature on a particular intervention, then the average effect size of that intervention is over-estimated by the available studies in that literature. This is not a failing of effect sizes but a bias that applies to the available literature. The result is that the scientific literature over-estimates the effects of an intervention. The statistical literature contains a considerable body of methods to use effect sizes to detect (and sometimes to adjust for) potential publication bias (Rothstein, Sutton, and Borenstein 2005; Shadish, Hedges, Pustejovsky, Boyajian, et al. 2014; Shadish, Hedges, and Pustejovsky 2014).

Publication bias analyses in meta-analysis are intended to do two things. The first is to assess the likelihood that publication bias may be present in the effect sizes being studied. Methods for doing this include some simple statistical tests (e.g., Begg and Mazumdar 1994; Egger, Davey Smith, Schneider, and Minder 1997); inspecting a funnel plot; and in meta-analyses that contain both published and unpublished studies, seeing if their effect sizes differ in the direction predicted by publication bias expectations. The second is to estimate what the average effect size would be if all

[35] Shadish, Zelinsky, Vevea, and Kratochwill (2014) have just completed a survey of several hundred SCD researchers on factors that influence decisions to publish. Between 5-15 percent of respondents said they were likely or very likely to drop a case prior to submitting for publication. Results are currently being prepared for publication.

studies were included in the meta-analysis. Methods for doing this include trim-and-fill analysis (Duval 2005; Duval and Tweedie 2000a, 2000b) and selection model methods (Hedges and Vevea 1996; Vevea and Hedges 1995). None of these methods is perfect, but their use will encourage SCD meta-analysts to think about whether the problem needs attention. Shadish, Hedges, and Pustejovsky (2014) provide worked-through examples for computing publication bias analyses. They found modest evidence that publication bias may exist in the Pivotal Response Training literature and that adjusting for that bias might result in lowering the effect size by about 25 percent.

Of course, the computation of an effect size will not, in itself, stop the practices that lead to publication bias from occurring. Indeed, to the extent that some SCD researchers believe that demonstrating a visually compelling functional relationship (and implicitly, then, a large statistical effect size) should be an important consideration in publication decisions, those researchers may argue we should in principle continue such practices. Researchers with this belief may then continue to give priority to the publication of large visual effects in their own work or in the review process. Even so, others are likely to compute an effect size on such data, and the results will then shed light on the possible effects of such practices on conclusions about what works. So in sum, our recommendation is to report a between-case effect size independently of the findings of a visual analysis about the effects of a treatment. If the SCD researcher does not do it, it is likely that reviewers who are not SCD researchers (such as those in federal and other research clearinghouses) will do so anyway.

In some respects, the issue of publication bias in SCD research may prove to be one of the most intellectually interesting but also contentious sources of discussion related to the present paper. The dilemma is this: For many SCD researchers, there is an assumption that only studies documenting a functional relation warrant publication. Studies that do not demonstrate a functional relation are considered uninterpretable; for example, a negative result may not mean that the treatment failed but that the treatment was not adequately implemented or that the outcome was not measured with sufficient reliability or validity. Still, a negative result may mean the treatment does not work in at least some cases, and in principle, it is quite feasible to assess whether treatment implementation or measurement unreliability are problems in any given study. So a need exists to better define the professional standards for publishing negative effects and the process for documenting intervention ineffectiveness. In principle, knowledge of what does not work should have just as much a place in

evidence-based practice reviews as knowledge of what does work. We cannot resolve this dilemma in the present paper, but we hope the methods we propose will help researchers in and out of the SCD community to study the issue.

6.3 Issues That Arise When Combining Results From SCDs With Results From Between-Groups Experiments

In section 6.1, we gave an example of a WWC review that did not include a possibly relevant SCD study, and we said WWC excluded it for good reason, namely that the study did not meet the pilot WWC SCD standards because it did not have the minimum three observations per phase that the standards require. This example thus suggests that just because a between-case effect size allows the reviewer to compare and combine effect sizes from SCDs and group studies, it does not mean the reviewer should do so. In this section, we describe the issues that should be considered when making the decision about whether and how to include SCD studies in a review with between-group studies.

The issue of whether and how to include SCDs is a special case of the more general issue of combining results from studies using different methods. That issue has been controversial since the earliest meta-analyses (e.g., Smith and Glass 1977), when critics said that combining randomized and nonrandomized group studies was unwise. Even today, this issue does not have a widely agreed-upon answer. For many years, for example, the Cochrane Collaboration (www.cochrane.org) relied nearly exclusively on evidence from randomized experiments to the exclusion not just of SCDs but of most other between-groups studies that did not use random assignment.

This nearly exclusive reliance on randomized experiments is changing. A special issue of the journal *Research Synthesis Methods* focused on issues that arise when including nonrandomized studies in systematic reviews of interventions (Reeves and Wells 2013), all of which should apply to SCDs as much as to other experiments that do not use random assignment. One issue is that randomized experiments are known in theory to yield statistically unbiased effects, something most nonrandomized experiments cannot claim. Thus, the reviewer needs methods to understand the degree of bias that may exist in nonrandomized experiments. A second issue is confounding, as when the treatment and control group are composed of people with different characteristics related

to outcome, often called selection bias. How to assess confounders and their impact becomes important. A third issue is selective reporting of results, which not only includes the publication bias matter just discussed but also includes other problems like selective reporting of analyses within a study even when that study was published. Solutions include referring to pre-study protocols or institutional review board submissions to see if the executed study reported all the planned analyses or contacting original authors about the work. A fourth issue is how to report results from studies with different designs in a systematic review. The reviewer may, for example, report results separately for different designs or have a table showing whether the people, interventions, outcomes, or settings differed in a systematic way across each study design.

Consider how these issues might relate to SCDs. The first issue is bias. Like most nonrandomized experiments, SCDs cannot claim to yield the statistically unbiased effects of a randomized experiment (see section 2.3.1). How, then, do we understand the quantity and quality of any bias that might be present? Comparing results from SCDs and randomized experiments on the same topic is a start and should be done, but it is far from definitive because SCDs may be done with different people, interventions, outcomes, or settings that can also affect the size of effects. Bias checklists specific to SCDs are useful, listing possible sources of bias and response options for understanding the biases that might apply. Bias may also differ for different kinds of SCDs, like the four basic SCDs outlined in section 2.1; indeed, addressing these potential design-based differences is the point of the pilot WWC SCD standards. Even if biases do not exist, SCDs may provide data about particular kinds of people, interventions, outcomes, or settings that are generally not available in randomized experiments, where that information is germane to the evidence-based practice decision at issue. Reviewers may need to take that into account when deciding whether to include SCDs in the review.

The second issue is confounding. The usual conception of selection bias, that people in the treatment group are different from those in the control, does not apply because the case acts as its own control in an SCD. Temporal confounding sometimes occurs in other interrupted time series when the composition of the sample in the time series changes over time. That will not occur when a single person is the case in an SCD, though it could occur when the case is an aggregate like a classroom whose composition could change over time. The most common confound in any time series, including an SCD, is that some event other than treatment occurs at the same time as

treatment and is the actual cause of the observed results (Campbell called this the history threat to internal validity). It is precisely in order to reduce the plausibility of such bias that most standards for SCD require three separate demonstrations of the effect, as we discussed in section 2.2.3. The confound could also be a change in outcome recording procedures concurrent with a change in phases. This is very unlikely in SCDs, where great emphasis is placed on constant and reliable measurement over time, though the issue has not been sufficiently explored empirically to be confident that the emphasis in theory is well implemented in practice.

The third issue is selective reporting, that is, a tendency to report favorable results and under-report negative results within one study. This is clearly not specific to SCDs—indeed there is good evidence that selective reporting of outcome measures occurs in many kinds of designs (e.g., Dal-Re and Caplan 2014; Dwan et al. 2008). Yet it has not received adequate attention in the SCD literature. For example, an SCD researcher studying several cases might decide not to include one of those cases in a publication because the case did not demonstrate a functional relationship or could not reach a stable baseline. Or the researcher might drop an observation within a case because it displayed unusual behavior like being unusually discrepant from other observations during baseline—the latter would generally be considered very poor professional practice in the SCD community, but some evidence exists it may sometimes happen. We need to learn how often such decisions occur in SCD research, the reasons for the decision, and what consequences might ensue from it.

The fourth issue also concerns reporting. The most common practice in reviews of between-groups studies is probably to report results separately for different designs (e.g., separately for randomized experiments, regression discontinuity designs, and nonequivalent comparison group designs), rather than combining results into one effect size. This is probably a good starting point when including SCDs in evidence-based practice reviews (e.g., reporting results from SCDs separately from the various kinds of between-groups designs above), at least until bias, confounding, and selective reporting are all better understood. Another option is sensitivity analysis, to see how the results of the review would change with and without the SCDs in the final summary. SCDs will probably also need different formats for tables that summarize individual studies, for example, reporting number of time points, number of cases, and kind of design, in addition to the usual practice of describing outcome measures, case characteristics, and setting. Some mechanism would need to be developed

for reporting the relatively rare cases in which random assignment is used to assign conditions to time, perhaps a footnote if nothing else (Kratochwill and Levin 2010). Finally, reporting should make particular note of the kinds of people, interventions, outcomes, or settings that are uniquely represented in SCDs compared to studies with other designs. Such information helps inform conclusions about generalizability. For example, it might be that the kinds of interventions studied in SCDs, typically with intense provider involvement with an individual case over a period of time, is systematically different from the kinds of interventions studied in group experiments.

All of these issues need discussion specific to SCDs, and full discussion with plausible solutions will take years. This paper can only raise the issues, not resolve them. Our own beliefs are optimistic on all counts and support the inclusion of SCDs in evidence-based practice reviews.

Computationally, the procedures for combining SCDs and between-groups experiments are the same as presented above for combining SCD effect sizes. Worked-through examples are not any different. However, we recommend that that the researcher code whether each effect size came from an SCD or some other design (or even more specific codes for the kind of SCD or the kind of between-groups design). Those codes could then be used as moderator variables to see whether different designs yield different effect size magnitudes when applied to the same question. If effect sizes from SCDs do eventually prove to be larger than those from between-groups designs, as suggested in section 3.1.2, the research community will need to address the wisdom or even the meaning of a pooled effect size over all designs. Shadish, Hedges, and Pustejovsky (2014) show worked-through examples of how to analyze moderator variables, so the researcher only needs to substitute the name(s) of their design variable(s) for the moderators used in those examples.

The fact that these between-case effect sizes are design comparable also allows their use to study differences between diverse experimental designs. For example, for several decades now between-groups researchers have been interested in what have come to be called within-study comparisons (WSCs; Cook, Shadish, and Wong 2008; Shadish and Cook 2009), studies that compare results from randomized and nonrandomized experiments in order to learn about such things as the conditions under which the latter can approximate the former and any systematic differences between design types that might make their direct comparison more complex. Between-case effect sizes allow us to extend that work to the comparison of SCDs to other experimental methodologies. The WSC

literature has burgeoned rapidly, but extensions of it to any form of time series is mostly lacking, so adding the study of SCDs to this mix would fill an important gap.

6.4 Design Standards as Inclusion Criteria in Reviews With SCDs

One key tenet of systematic reviews is to have explicit inclusion and exclusion criteria for what studies are in the review. One class of such criteria is design standards, the idea that studies should have used a design that is appropriate to the question being asked and that meets professionally defined criteria for methodological quality. The WWC *Procedures and Standards Handbook* (U.S. Department of Education 2014a), for example, specifies designs that meet standards with or without reservations and does so for such diverse designs as randomized experiments, regression discontinuity design, and SCDs (as we mentioned before, these are pilot standards for SCDs). For randomized experiments, for example, the combination of overall or differential attrition from conditions after random assignment can change the rated quality of the study or remove it from the review entirely. Even with attrition, studies can qualify for inclusion if the researcher demonstrates baseline equivalence between remaining treatment and control participants. This leads, of course, to debate about what constitutes acceptable equivalence, and such standards can change as more is learned about the impact of any particular criterion on the bias and precision of the resulting effect size estimate.

For SCDs, the criteria that WWC uses to define whether a study meets standards with or without reservations were listed in section 2.2.3 of this paper. Many reviewers will use such standards as they plan a review that might include SCD studies. That being said, we should remember that professional consensus about standards can change with experience and empirical data. For example, single-case researchers have been encouraged to identify a practice as evidence-based using the 5-3-20 rule, which states that sufficient evidence exists to assess an intervention as evidence based if (1) at least 5 SCD studies are available that each meet both design and evidence standards; (2) those studies are conducted by at least three different research teams in different locations; and (3) the combined number of cases across all qualified studies is at least 20. Just like attrition and equivalence thresholds for randomized experiments can change, so too can the three specific numbers in the 5-3-20 rule change based on further analysis of empirical study. Imagine, for example, that substantial evidence accumulates that a 4-3-20 rule consistently gives the same average

effect size as a 5-3-20 rule. Conversely, imagine that evidence suggests that 20 cases are simply not enough to document a reliable effect. Standards may then change. Of course, the only way to accumulate such evidence is to study these matters empirically and report results publicly. Hence, we encourage interested scholars to pursue such research.

7. Future Directions and Development for Between-Case Single-Case Design Effect Sizes

The present paper suggests practices that the authors believe will improve analysis within SCDs, increase the chances that SCDs are included in formal meta-analyses, increase the use of SCDs to identify evidence-based practices, and increase the ability of single-case researchers to identify moderating variables that may guide ongoing lines of research by searching for variables across studies that influence the size of effect and that might not have sufficient power or variability to be detected with any one SCD study. This final chapter suggests the kinds of activities that are needed to foster continued development of between-case effect sizes.

7.1. Research to Improve Between-Case Effect Sizes

Despite their advantages for the inclusion of SCDs in the evidence-based practice literature, between-case effect sizes in SCDs are very new and will benefit from considerable additional development. This chapter discusses some of the directions and developments that are needed (many of which are already in progress).

Extensions to Other SCD Designs. Between-case effect sizes are developed for reversal and multiple baseline designs but not explicitly developed for other SCDs. Two extensions are salient. The first is to other basic designs such as alternating treatment or changing criterion designs, which will require additional basic statistical research. The second is demonstrating how to apply effect sizes to complicated designs, such as when a time series starts with a baseline and then switches to an alternating treatments design to test two interventions, followed by a generalization phase.

In principle, the answer to this second extension is less complex than might first seem. Most effect sizes used in evidence-based practice reviews, including reviews of between-groups designs, are computed on the comparison of just two conditions at a time. For example, a randomized experiment may have three conditions (two treatments and a control) and use a factorial design in which those conditions are crossed with gender (male, female), resulting in a 3 x 2 factorial design. The main analysis might be a regression, including main effects and interactions, and it is certainly

possible to compute an omnibus effect size that incorporates all of these conditions at once (e.g., a measure of percentage of variance accounted for by the factors and conditions in total). However, few between-groups researchers would compute the omnibus effect size because it is usually of little theoretical or practical interest. A little thought suggests the same may be largely true of SCDs. For the complex design described in the previous paragraph, an omnibus effect size would likely have little meaning. Instead, just like the between-groups researcher, the SCD researcher would likely be more interested in various pairwise comparisons—e.g., the effect of baseline compared to Treatment A or of Treatment B to generalization. These pairwise comparisons are easily done with existing methods when the complex design is composed of multiple baseline or reversal designs (though an adjustment for doing multiple tests may be warranted if conducting multiple tests of statistical significance). For instance, comparing baseline to Treatment A simply involves selecting only the outcome data for those two conditions and then computing either the multiple baseline or reversal effect size, depending on which fits the data.

All this is not to say that omnibus tests are of no interest. For example, Shadish, Kyse, and Rindskopf (2013; see also Shadish, Zuur, and Sullivan 2014) show examples of how multilevel models can be used to analyze all the data in a complex SCD at once. Such models may come closer to reflecting the complexities that SCD researchers use in making judgments about the existence of functional relationships. However, they do not serve the same function as a between-case effect size. One reason is that, in their usual unstandardized form, they typically do not place diverse outcomes into a common metric. Another reason is that no statistical proofs currently exist that the effect estimate they yield is identical to the usual between-case effect size. Finally, such models usually summarize treatment effects over all features of the design rather than on a given pair-wise comparison that is the focus of a between-case effect size. For example, a multilevel analysis of an SCD study that includes baseline, treatment, generalization, and follow-up phases may produce an effect over all phases, but a between-case effect size always focuses on a comparison between just two phases. These limitations are remediable. Still, it is unlikely that any single statistic will meet all the needs of SCD researchers, just as is the case with between-groups researchers.

Extensions Regarding Trend. One need is for continued work on incorporating trend in an accessible way. The Pustejovsky et al. (2014) effect size makes significant progress on this but may not be as accessible to SCD researchers as a method in a commonly used program like SPSS or with

a graphical user interface. Further, little thought has yet been given to the nature of between-case effect sizes when nonlinear trend may be present; methods for including nonlinearities in the analysis of SCD data exist (e.g., Shadish, Zuur, and Sullivan, 2014; Sullivan, Shadish, and Steiner in press), but extending them to include the estimation of a between-case effect size remains to be done. For both linear and nonlinear trend, more needs to be done to address the point at which treatment effects are assessed when trend is present. Work on the latter could also address an issue important to SCD researchers: the immediacy of the change after the introduction of treatment. When the researcher does have an a priori hypothesis about when change should occur, effect sizes that incorporate trend can assess effects at or around that specific time. Without a prior hypothesis, more exploratory methods will be useful (e.g., Shadish, Kyse, and Rindskopf 2013; Shadish, Zuur, and Sullivan 2014), though they may not result in a between-case effect size at their current stage of development.

Related work would concern how much trend is too much for accurately estimating treatment effects. For example, too much trend might be defined as statistically significant trend, though power concerns for testing the presence of trend in samples typical of SCD research are nontrivial. One could also review existing literatures to tell whether taking into account trend of any kind (e.g., linear, nonlinear) in the statistical analysis yields quantitatively or qualitatively different conclusions about treatment effects.

When The Predicted Effect Is Change in Slope or in Variability. The preceding section assumed the effect of interest was a change in level, with trend in a form that made it more difficult to detect that change. Sometimes, however, the SCD researcher will predict (or find retrospectively) that a change in trend is the effect of an intervention. In principle, once solutions are developed for the simple trend problem, predicted changes in trend should be easy to incorporate into the effect size, mostly through specification of the time point within the treatment phase at which the researcher wishes to measure that change in trend. This challenge is more significant with data sets demonstrating strong ceiling or floor effects where the slope and variability are truncated or where the researcher predicts a change only in variability of the outcome measure as a result of the intervention. Further work on these latter problems would be worthwhile.

Extensions to Other Outcome Metrics. Reason exists to think current between-case effect sizes may be at least somewhat robust to violations of the assumption of normality of residuals. Still, given the prevalence of patently non-normal outcome variables in SCD research, relying on the possibility of robustness is less reassuring than (a) empirical investigations into the robustness properties of estimators that use the normality assumption and (b) effect size estimators developed specifically for outcomes with non-normally distributed errors. Examples include Poisson distributions for count data and binomial distributions for rate data. Even the latter are not a complete set of options, however. Some evidence is emerging that count and rate data in SCDs are likely to be overdispersed, that is, have more variability than is expected given the properties of the distribution (Shadish, Kyse, and Rindskopf 2013; Shadish, Zuur, and Sullivan 2014). So the research possibilities here are extensive.

Number of Cases. Another limitation is the requirement to have multiple cases in order to compute a between-case effect size. In part, this requirement is definitional—one cannot have a between-case effect size without having multiple cases. That does not mean, however, that no progress could be made to address the problem. It may be, for example, that one could compute the standard deviation in equation (2) for $d_{between}$, and then substitute that for the denominator of equation (1) for d_{within} to yield an effect size for each case. The viability and statistical properties of such an approach require further research. Similarly, it may be that current or new within-case effect sizes could be developed in a way that shows how to convert them to between-case effect sizes. The latter is not as simple as, say, averaging all the overlap statistics in a study, however. That average would still be based on within-case standardization and so not comparable to between-groups effect sizes.

What should an evidence-based practice reviewer do when it is not possible to compute a between-case effect size on studies that have only one or two cases? Although evidence suggests studies with so few cases are the minority (Shadish and Sullivan 2011), not using even a few studies wastes information. Until future technical developments solve this problem, the reviewer can do two things. One is to review them qualitatively in the same review as the studies on which a between-case effect size can be computed, focusing especially on whether they differ from other studies in important ways—do they use systematically different kinds of cases, treatments, outcomes, or settings, for example? Second, pertinent statistics can still be computed. One case allows computing the numerator of the effect size, that is, the mean difference between conditions; that can be

compared to the numerators of studies with three cases for overall comparability, at least when the same outcome variable is used. Two cases allow computing the denominator of the effect size (one can't compute a variance between cases without at least two cases), which then allows computing the effect size itself for similar comparison. The researcher should be cautious about this option, in particular recognizing that such few cases may result in an effect size that is quite different in magnitude from the other effect sizes by chance alone. Three cases are required (for statistical reasons) to compute the sampling error of the effect size (and so to compute significance tests and confidence intervals and to compute the inverse variance weights for meta-analytic integration) and to compute the correction for small sample bias. So, much can be learned from studies with one or two cases even if all options are not possible.

The accumulation of empirical evidence about the magnitude of between-case versus within-case variability may also present research opportunities analogous to those that have arisen in between-groups studies with clustered sampling. Compendia of empirical information have been useful in providing guidelines for converting among effect sizes of different types (roughly analogous to $d_{between}$ and d_{within} in the SCD context). Such information might be as useful in informing design decisions in SCDs as it has been in the design of between-groups design.

Clarifying Relations of Within-Case Effect Sizes to Between-Case Effect Sizes. As currently developed, within-case and between-case effect sizes have no clear relationship to each other. If conditions could be specified under which the two variants could be equated, that would allow SCD researchers to reap the benefits of both—descriptions of effect size for each case and descriptions of effect size in a metric comparable to that used in between-groups studies.

7.2 Conclusions

Largely due to the impact of evidence-based practice, SCD research is at a historic crossroad. It can continue forward as it has for many decades, relying nearly exclusively on the methods that have served it so well, such as visual analysis, but that have not necessarily led SCDs to be included in many pertinent evidence-based practice reviews. Or it can take a risk and expand to consider whether at least some statistics can be useful to SCD work, especially in garnering the recognition SCD research needs to be included in evidence-based practice reviews. SCD researchers are in a very

unusual position in this regard. The value of research findings using SCDs is getting increasing recognition, and the credibility of SCD methodology to document experimental effects is solidifying. The next step for improving the accessibility of single-case research results is building agreement about effect size measures that open SCDs to a broader audience and broader forms of analysis.

References

Allison, D. B., and Gorman, B. S. (1993). Calculating Effect Sizes for Meta-Analysis: The Case of the Single Case. *Behaviour Research and Therapy, 31*, 621-631.

American Psychological Association. (2002). Ethical Principles of Psychologists and Code of Conduct. *American Psychologist, 57*: 1060-1073.

American Speech-Language-Hearing Association. (2004). *Evidence-Based Practice in Communication Disorders: An Introduction* [Technical Report]. Available from http://shar.es/11yOzJ or http://www.asha.org/policy/TR2004-00001/.

Arthur, W., Bennett, W., and Huffcutt, A.I. (2001). *Conducting Meta-Analysis Using SAS*. New York, NY: Psychology Press.

Baer, D.M. (1977). Perhaps It Would Be Better Not to Know Everything. *Journal of Applied Behavior Analysis, 10*: 167-172.

Barlow, D., Nock, M., and Hersen, M. (2009). *Single Case Experimental Designs: Strategies for Studying Behavior for Change* (3rd Ed.). Boston, Massachusetts: Allyn and Bacon.

Begg, C.B., and Mazumdar, M. (1994). Operating Characteristics of a Rank Correlation Test for Publication Bias. *Biometrics, 50*: 1088-1101.

Beretvas, S.N., and Chung, H. (2008). An Evaluation of Modified R2-Change Effect Size Indices for Single-Subject Experimental Designs. *Evidence-Based Communication Assessment and Intervention, 2*: 120-128. doi: 10.1080/17489530802446328.

Borenstein, M., Hedges, L.V., Higgins, J.P.T., and Rothstein, H.R. (2005). *Comprehensive Meta-Analysis, Version 2*. Englewood, NJ: Biostat.

Borenstein, M., Hedges, L.V., Higgins, J.P.T., and Rothstein, H.R. (2009). *Introduction to Meta-Analysis*. West Sussex, UK: Wiley.

Brossart, D. F., Parker, R. I., Olson, E. A., and Mahadevan, L. (2006). The Relationship Between Visual Analysis and Five Statistical Analyses in a Simple AB Single-Case Research Design. *Behavior Modification, 30*, 531-563.

Bulté, I., and Onghena, P. (2012). When the Truth Hits You Between the Eyes: A Software Tool for the Visual Analysis of Single-Case Experimental Data. *Methodology, 8*, 104-114.

Campbell, D.T. (1957). Factors Relevant to the Validity of Experiments in Social Settings. *Psychological Bulletin, 54*: 297-312.

Campbell, D.T., and Stanley, J.C. (1966). *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand-McNally.

Campbell, J. M. (2004). Statistical Comparison of Four Effect Sizes for Single-Subject Designs. *Behavior Modification*, 28, 234-246.

Cicerone, K., Dahlberg, C., Kalmar, K., Langenbahn, D., Malec, J., Bergquist, T., Felicetti, T., Giacino, J., Harley, J., Harrington, D., Herzog, J., Kneipp, S., Laatsch, L., and Morse P. (2000). Evidence-Based Cognitive Rehabilitation: Recommendations for Clinical Practice. *Archives of Physical Medicine and Rehabilitation, 81*: 1596-1615.

Cicerone, K., Dahlberg, C., Malec, J.F., Langenbahn, D.M., Thomas, F., Kneipp, S., Ellmo, W., Kalmar, K., Giacino, J., Harley, J., Laatsch, L., Morse, P., and Catanese, J. (2005). Evidence-Based Cognitive Rehabilitation: Updated Review of the Literature From 1998 Through 2002. *Archives of Physical Medicine and Rehabilitation, 86*: 1681-1692.

Cicerone, K.D., Langenbahn, D.M., Braden, C., Malec, J.F., Kalmar, K., Fraas, M., Felicetti, T., Laatsch, L., Harley, J.P., Bergquist, T., Azulay, J., Cantor, J., and Ashman, T. (2011). Evidence-Based Cognitive Rehabilitation: Updated Review of the Literature From 2003 Through 2008. *Archives of Physical Medicine and Rehabilitation, 92*: 519-530.

Cochran, W.G. (1965) The Planning of Observational Studies of Human Populations (with Discussion). *Journal of the Royal Statistical Society (Series A)*, *128*: 134-155.

Codding, R.S., Livanis, A., Pace, G.M., and Vaca, L. (2008). Using Performance Feedback to Improve Treatment Integrity of Classwide Behavior Plans: An Investigation of Observer Reactivity. *Journal of Applied Behavior Analysis*, *41*(3): 417-422. doi: 10.1901/jaba.2008.41-417.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Conelea, C.A., and Woods, D.W. (2008). Examining the Impact of Distraction on Tic Suppression in Children and Adolescents With Tourette syndrome. *Behaviour Research and Therapy*, *46*: 1193-1200. doi:10.1016/j.brat.2008.07.005

Cook, B.G., Buysse, V., Klingner, J., Landrum, T.J., McWilliam, R.A., Tankersley, M., and Test, D. W. (2014). CEC's Standards for Classifying the Evidence Base of Practices in Special Education. *Remedial and Special Education, 39*: 305-318. doi: 0741932514557271.

Cook, T.D. (1990). The Generalization of Causal Connections: Multiple Theories in Search of Clear Practice. In L. Sechrest, E. Perrin, and J. Bunker (Eds), *Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data* (pp. 9-31) (DHHS Publication No. (PHS) 90-3454). Rockville MD: Department of Health and Human Services.

Cook, T.D., and Campbell, D.T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Chicago: Rand-McNally.

Cook, T.D., Cooper, H.M., Cordray, D.S., Hartmann, H., Hedges, L.V., Light, R.J., Louis, T.A., and Mosteller, F. (Eds.). (1991). *Meta-Analysis for Explanation: A Casebook*. New York: Russell Sage Foundation.

Cook, T.D., Shadish, W.R., and Wong, V.C. (2008). Three Conditions Under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings From Within-Study Comparisons. *Journal of Policy Analysis and Management, 27*: 724-750.

Council for Exceptional Children. (2014). Council for Exceptional Children: Standards for Evidence-Based Practices in Special Education. *Exceptional Children, 80*: 504-511.

Cronbach, L.J. (1982). *Designing Evaluations of Educational and Social Programs*. San Francisco: Jossey-Bass.

Dal-Re, R., and Caplan, A.L. (2014). Time to Ensure That Clinical Trial Appropriate Results Are Actually Published. *European Journal of Clinical Pharmacology, 70*: 491-493.

Duval, S. J., and Tweedie, R. L. (2000a). Trim and Fill: A Simple Funnel-Plot-Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis. *Biometrics, 56*, 455-463.

Duval, S. J., and Tweedie, R. L. (2000b). A Nonparametric "Trim and Fill" Method of Accounting for Publication Bias in Meta-Analysis. *Journal of the American Statistical Association, 95(449),* 89-98.

Duval, S. J. (2005). The Trim and Fill Method. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication Bias in Meta-Analysis: Prevention, Assessment, and Adjustments* (pp. 127–144). Chichester, England: Wiley.

Dwan K., Altman D.G., Arnaiz J.A., Bloom J., Chan A-W., Cronin, E., Decullier, E., Easterbrook, P.J., Von Elm, E., Gamble, C., Ghersi, D., Ioannidis, J.P., Simes, J., and Williamson, P.R. (2008). Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias. *PLoS ONE*, *3*: e3081. doi:10.1371/journal.pone.0003081

Egger, M., Davey Smith, G., Schneider, M., and Minder, C. (1997). Bias in Meta-Analysis Detected by a Simple, Graphical Test. *British Medical Journal, 315*: 629-634.

Fabiano, G.A., Pelham, W.E., Coles, E.K., Gnagy, E.M., Chronis-Tuscano, A., and O'Connor, B.C. (2009). A Meta-Analysis of Behavioral Treatments for Attention-Deficit/Hyperactivity Disorder. *Clinical Psychology Review, 29*: 129-140.

Field, A.P., and Gillett, R. (2010). How to Do a Meta-Analysis. *British Journal of Mathematical and Statistical Psychology, 63*: 665–694. http://dx.doi.org/10.1348/000711010X502733.

Gabler, N.B., Duan, N., Vohra, S., and Kravitz, R.L. (2011). N-of-1 Trials in the Medical Literature: A Systematic Review. *Medical Care, 49*: 761-768.

Gast, D.L., and Ledford, J.R. (2014). *Single Case Research Methodology: Applications in Special Education and Behavioral Sciences* (2nd ed.). New York: Routledge.

Glass, G.V. (1976) Primary, Secondary and Meta-Analysis of Research. *Educational Researcher*, *5*: 3-8

Glover, A.C., Roane, H.S., Kadey, H.J., and Grow, L.L. (2008). Preference for Reinforcers Under Progressive- and Fixed-Ration Schedules: A Comparison of Single and Concurrent Arrangements. *Journal of Applied Behavior Analysis*, *41*(2): 163-176. doi: 10.1901/jaba.2008.41-163.

Gurka, M.J., Edwards, L.J., and Muller, K.E. (2011). Avoiding Bias in Mixed Model Inference for Fixed Effects. *Statistics in Medicine, 30*: 2696–2707. doi:10.1002/sim.4293

Hedges, LV., and Olkin, I. (1985). *Statistical Methods for Meta Analysis*. Orlando, FL: Academic Press.

Hedges, L.V., and Pigott, T.D. (2001). The Power of Statistical Tests in Meta-Analysis. *Psychological Methods, 6*: 203-217.

Hedges, L.V., and Pigott, T.D. (2004). The Power of Statistical Tests for Moderators in Meta-Analysis. *Psychological Methods*, *9*: 426-445.

Hedges, L.V., Pustejovsky, J., and Shadish, W.R. (2012). A Standardized Mean Difference Effect Size for Single-Case Designs. *Research Synthesis Methods*, *3*: 224-239.

Hedges, L.V., Pustejovsky, J., and Shadish, W.R. (2013). A Standardized Mean Difference Effect Size for Multiple Baseline Designs Across Individuals. *Research Synthesis Methods*, *4*: 324-341. doi: 10.1002/jrsm.1086.

Hedges, L.V., and Shadish, W.R. (2015a). Power Analysis for Single Case Designs. Unpublished manuscript available from the authors.

Hedges, L.V., and Shadish, W.R. (2015b). Power Analysis for Multiple Baseline Designs. Unpublished manuscript available from the authors.

Hedges, L.V. and Vevea, J. L. (1996). Estimating Effect Size Under Publication Bias: Small Sample Properties and Robustness of a Random Effects Selection Model. *Journal of Educational and Behavioral Statistics, 21*, 299-333.

Himle, M.B., Woods, D.W., and Bunaciu, L. (2008). Evaluating the Role of Contingency in Differentially Reinforced Tic Suppression. *Journal of Applied Behavior Analysis*, *41*(2): 285-289. doi:10.1901/jaba.2008.41-285

Hitchcock, J.H., Horner, R.H., Kratochwill, T.R., Levin, J.R., Odom, S.L., Rindskopf, D.M., and Shadish, W.R. (2014). The What Works Clearinghouse Single-Case Design Pilot Standards: Who Will Guard the Guards? *Remedial and Special Education, 35*: 145-152. doi: 10.1177/0741932513518979.

Horner, R.D., and Baer, D.M. (1978). Multiple-Probe Technique: A Variation on the Multiple Baseline. *Journal of Applied Behavior Analysis, 11*: 189-196.

Horner, R.H., Carr, E.G., Halle, J., McGee, G., Odom, S., and Wolery, M. (2005). The Use of Single Subject Research to Identify Evidence-Based Practice in Special Education. *Exceptional Children*, *71*: 165-179.

Horner, R.H., and Odom, S.L. (2014). Constructing Single-Case Research Designs: Logic and Options. In T. Kratchowill (Ed.), *Single-Case Intervention Research: Methodological and Data-Analysis Advances* (pp. 27-52). Washington D.C.: American Psychological Association.

Howick, J., Chalmers, I., Glasziou, P., Greenhaigh, T., Heneghan, C., Liberati, A., et al. (2011). The 2011 Oxford CEBM Evidence Table (Introductory Document). *Oxford Centre for Evidence-Based Medicine*. http://www.cebm.net/index.aspx?o=5653

IBM Corp. (2013). *IBM SPSS Statistics for Windows, Version 22.0*. Armonk, NY: IBM Corp.

Imbens, G.W., and Rubin, D.B. (in press). *Causal Inference in Statistics, and in the Social and Biomedical Sciences*. New York: Cambridge University Press.

Ioannidis, J.P.A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*: e124.

Ioannidis, J.P.A. (2012). Why Science Is Not Necessarily Self-Correcting. *Perspectives on Psychological Science, 7*: 645–654. doi: 10.1177/1745691612464056.

Kazdin, A. (1982). *Single-Case Research Designs: Methods for Clinical and Applied Settings*. Oxford University Press, New York.

Kazdin, A.E. (2011). *Single-Case Research Designs: Methods for Clinical and Applied Settings* (2nd ed.). New York: Oxford University Press.

Kratochwill, T.R., Hitchcock, J., Horner, R.H., Levin, J.R., Odom, S.L., Rindskopf, D.M., and Shadish, W.R. (2010). Single-Case Designs Technical Documentation. Available from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Kratochwill, T.R., Hitchcock, J., Horner, R.H., Levin, J.R., Odom, S.L., Rindskopf, D.M, and Shadish, W.R. (2013). Single-Case Intervention Research Design Standards. *Remedial and Special Education, 34*: 26-38. doi: 10.1177/0741932512452794.

Kratochwill, T.R., and Levin, J.R. (2010). Enhancing the Scientific Credibility of Single-Case Intervention Research: Randomization to the Rescue. *Psychological Methods*, *15*: 122-144.

Kratochwill, T.R., and Levin, J.R. (Eds.). (2014). *Single-Case Intervention Research: Methodological and Statistical Advances*. Washington, D.C.: American Psychological Association.

Kratochwill, T.R., and 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.

Lambert, M.C., Cartledge, G., Heward, W.L., and Lo, Y. (2006). Effects of Response Cards on Disruptive Behavior and Academic Responding During Math Lessons by Fourth-Grade Urban Students. *Journal of Positive Behavior Interventions, 8e*: 88-99.

Lau, J., Antman, E.M., Jimenez-Silva, J., Kupelnick, B., Mosteller, F., and Chalmers, T.C. (1992). Cumulative Meta-Analysis of Therapeutic Trials for Myocardial Infarction. *The New England Journal of Medicine, 23*: 248-254.

LeBel, E.P., Borsboom, D., Giner-Sorolla, R., Hasselman, R., Peters, K.R., Ratliff, K.A., and Smith, C.T. (2013). PsychDisclosure.org: Grassroots Support for Reforming Reporting Standards in Psychology. *Perspectives on Psychological Science, 8*: 424-432.

Lipsey, M.W., and Hurley, S.M. (2009). Design Sensitivity: Statistical Power for Applied Experimental Research. In L. Bickman and D. Rog (Eds). *The SAGE Handbook of Applied Social Research Methods* (2nd ed., pp. 44-76). Thousand Oaks, California: Sage Publications.

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). The BUGS Project: Evolution, Critique and Future Directions. *Statistics in Medicine, 28*: 3049-3067. doi:10.1002/sim.3680. PMID 19630097

Maggin, D.M., O'Keeffe, B.V., and Johnson, A.H. (2011). A Quantitative Synthesis of Methodology in the Meta-Analysis of Single-Subject Research for Students With Disabilities: 1985–2009. *Exceptionality: A Special Education Journal, 19*: 109-135. doi: 10.1080/09362835.2011.565725.

Mahoney, M. (1977). Publication Prejudices: An Experimental Study of Confirmatory Bias on the Peer Review System. *Cognitive Therapy and Research, 1*: 161-175.

Manolov, R., Arnau, J., Solanas, A., and Bono, R. (2010). Regression-Based Techniques for Statistical Decision Making in Single-Case Designs. *Psicothema, 22*, 1026-1032.

Manolov, R., Gast, D. L., Perdices, M., and Evans, J. J. (2014). Single-Case Experimental Designs: Reflections on Conduct and Analysis. *Neuropsychological Rehabilitation, 24*, 634-660.

Manolov, R., and Solanas, A. (2008). Comparing N = 1 Effect Size Indices in Presence of Autocorrelation. *Behavior Modification*, *32*, 860-875.

Marso, D., and Shadish, W.R. (2014). *User Guide for DHPS, D_Power, and GPHDPwr SPSS Macros* (Version 1.0). Unpublished manual available from http://faculty.ucmerced.edu/wshadish/software/software-meta-analysis-single-case-design.

Moeyaert, M., Ferron, J.M., Beretvas, S.N., and 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, ISSN 0022-4405, http://dx.doi.org/10.1016/j.jsp.2013.11.003.

Moher, D., Hopewell, S., Schulz, K.F., Montorid, V. Gøtzsche, P.C., Devereaux, P.J., Elbourne, D., Egger, M., and Altman, D.G. (2010). CONSORT 2010 Explanation and Elaboration: Updated Guidelines for Reporting Parallel Group Randomised Trials. *Journal of Clinical Epidemiology, 63*: e1-e37.

Morgan, S.L., and Winship C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge: Cambridge University Press.

NIMH Collaborative Data Synthesis for Adolescent Depression Trials Study Team (2013). Advancing Science Through Collaborative Data Sharing and Synthesis. *Perspectives on Psychological Science, 8*: 433-444. doi: 10.1177/1745691613491579.

Neyman, J. (1923). *Sur les Applications de la Theorie des Probabilites aux Experiences Agricoles: Essai des Principes*. Master's Thesis (1923). Excerpts reprinted in English, *Statistical Science, 5*: 463-472. (D.M. Dabrowska, and T. Speed, Translators.)

Odom, S.L., and Lane, K.L. (2014). The Applied Science of Special Education: Quantitative Approaches, the Questions They Address, and How They Inform Practice. In L. Florian (Ed.), *The SAGE Handbook of Special Education* (2nd ed., pp. 369-388). Thousand Oaks, CA: SAGE.

Open Science Collaboration. (2012). An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science. *Perspectives on Psychological Science, 7*: 657-660.

Owens, S.H., Fredrick, L.D., and Shippen, M.E. (2004). Training a Paraprofessional to Implement "Spelling Mastery" and Examining its Effectiveness for Students With Learning Disabilities. *Journal of Direct Instruction, 4*(2): 153-172.

Perone, M. (1999). Statistical Inference in Behavior Analysis: Experimental Control Is Better. *The Behavior Analyst, 22*: 109-116.

Pustejovsky, J.E. (2014*). scdhlm: Estimating Hierarchical Linear Models for Single-Case Designs*. R Package Version 0.2.0. Available for download at http://blogs.edb.utexas.edu/pusto/software/

Pustejovsky, J.E., Hedges, L.V., and Shadish, W.R. (2014). Design-Comparable Effect Sizes in Multiple Baseline Designs: A General Modeling Framework. *Journal of Educational and Behavioral Statistics, 39*: 368–393. doi:10.3102/1076998614547577.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org/

Rao, P.A., Beidel, D.C., and Murray, M.J., (2008). Social Skills Interventions for Children With Asperger's Syndrome or High-Functioning Autism: A Review and Recommendations. *Journal of Autism and Developmental Disorders, 38*: 353-361.

Reeves, B.C., and Wells, B.A. (2013). Inclusion of Non-Randomized Studies in Systematic Reviews [Special Issue]. *Research Synthesis Methods, 4*(1).

Reichardt, C.S. (2006). The Principle of Parallelism in the Design of Studies to Estimate Treatment Effects. *Psychological Methods, 11*: 1-18.

Reichow, B., Volkmar, F., and Cicchetti, D. (2008). Development of the Evaluative Method for Evaluating and Determining Evidence-Based Practices in Autism. *Journal of Autism and Developmental Disorders, 38*, 1311-1319.

Reynolds, K.D., and West, S.G., (1987). A Multiplist Strategy for Strengthening Nonequivalent Control Group Designs. *Evaluation Review*, *11*, 691-71

Rogers, H., and Swaminathan, H. (2009). SINGSUB (1.0) [Computer program]. Unpublished computer program.

Rogers, L.A., and Graham, S. (2008). A Meta-Analysis of Single Subject Design Writing Intervention Research. *Journal of Educational Psychology*, *100*(4): 879-906.

Rosenbaum, P.R. (2002). *Design of Observational Studies* (2nd ed). New York: Springer.

Rothstein, H., Sutton, A J., and Borenstein, M. (Eds.). (2005). *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. New York, NY: Wiley.

Rubin, D.B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology, 66*: 688-701.

Salsburg, D. (2001). *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. New York, NY: Henry Holt.

Schmidt, F.L., and Hunter, J.E. 2015. *Methods of Meta-Analysis: Correction Error and Bias and Research Findings*. Sage Publications, Thousand Oaks, California.

Schreibman, L., Stahmer, A.C., Barlett, V.C., and Dufek, S. (2009). Brief Report: Toward Refinement of a Predictive Behavioral Profile for Treatment Outcome in Children With Autism. *Research in Autism Spectrum Disorders*, *3*: 163-172.

Schutte, N.S., Malouff, J.M., and Brown, R.F. (2008). Efficacy of an Emotion-Focused Treatment for Prolonged Fatigue. *Behavior Modification, 32*: 699-713. doi:10.1177/0145445508317133.

Shadish, W.R. (1995). The Logic of Generalization: Five Principles Common to Experiments and Ethnographies. *American Journal of Community Psychology, 23*: 419-428.

Shadish, W.R. (2010). Campbell and Rubin: A Primer and Comparison of Their Approaches to Causal Inference in Field Settings. *Psychological Methods, 15*: 3-17.

Shadish, W.R. (ed.). (2014a). Analysis and Meta-Analysis of Single-Case Designs [Special issue]. *Journal of School Psychology, 52*(2).

Shadish, W.R. (2014b). Statistical Analyses of Single-Case Designs: The Shape of Things to Come. *Current Directions in Psychological Science*, *23*: 139-146. doi: 10.1177/0963721414524773.

Shadish, W.R., Brasil, I.C.C., Illingworth, D.A., White, K., Galindo, R., Nagler, E.D., and Rindskopf, D.M. (2009). Using UnGraph® to Extract Data From Image Files: Verification of Reliability and Validity. *Behavior Research Methods, 41*: 177-183.

Shadish, W.R., Cook, T.D., and Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin.

Shadish, W.R., and Haddock, C.K. (2009). Combining Estimates of Effect Size. In H.M. Cooper, L.V. Hedges, and J.C. Valentine (Eds.), The Handbook of Research Synthesis and Meta-Analysis (2nd ed., pp. 257-277). New York: Russell Sage Foundation.

Shadish, W.R., Hedges, L.V., and Pustejovsky, J.E. (2014). Analysis and Meta-Analysis of Single-Case Designs With a Standardized Mean Difference Statistic: A Primer and Applications. *Journal of School Psychology, 52*: 123-147.

Shadish, W.R., Hedges, L.V., Pustejovsky, J.E., Boyajian, J.G., Sullivan, K.J., Andrade, A., and Barrientos, J.L. (2014). A *d*-Statistic for Single-Case Designs That Is Equivalent to the Usual Between-Groups *d*-Statistic. *Neuropsychological Rehabilitation, 24*: 528-553. http://dx.doi.org/10.1080/09602011.2013.819021

Shadish, W.R., Kyse, E.N., and Rindskopf, D.M. (2013). Analyzing Data From Single-Case Designs Using Multilevel Models: New Applications and Some Agenda Items for Future Research. *Psychological Methods*, *18*: 385-405. doi: 10.1037/a0032964.

Shadish, W.R., and Lecy, J.D. (2015). The Meta-Analytic Big Bang. *Research Synthesis Methods*. doi: 10.1002/jrsm.1132.

Shadish, W.R., and Rindskopf, D.M. (2007). Methods for Evidence-Based Practice: Quantitative Synthesis of Single-Subject Designs. In G. Julnes and D.J. Rog (Eds.), *Informing Federal Policies on Evaluation Method: Building the Evidence Base for Method Choice in Government Sponsored Evaluation* (pp. 95-109). San Francisco: Jossey-Bass.

Shadish, W.R., and Sullivan, K.J. (2011). Characteristics of Single-Case Designs Used to Assess Intervention Effects in 2008. *Behavior Research Methods, 43*: 971-980. doi 10.3758/s13428-011-0111-y. ERIC #ED530280.

Shadish, W.R., and Zuur, A.F. (2014, October). *Power Analysis for Negative Binomial Glmms for Single-Case Designs*. Society for Multivariate Experimental Psychology, Nashville, TN.

Shadish, W.R., Zuur, A.F., and Sullivan, K.J. (2014). Using Generalized Additive (Mixed) Models to Analyze Single Case Designs. *Journal of School Psychology*, *52*(2): 149-178, ISSN 0022-4405, http://dx.doi.org/10.1016/j.jsp.2013.11.004.

Sham, E., and Smith, T. (2014). Publication Bias in Studies of an Applied Behavior Analytic Intervention: An Initial Analysis. *Journal of Applied Behavior Analysis, 47*: 663-678.

Shamseer, L., Sampson, M., Bukutu, C., Nikles, J., Tate, R., Johnson, B. C., Zucker, D. R., Shadish, W., Kravitz, R., Guyatt, G., Altman, D.G., Moher, D., Vohra, S., and the CENT Group. (2015). CONSORT extension for N-of-1 Trials (CENT) 2015: Explanation and Elaboration. *British Medical Journal*. doi: http://dx.doi.org/10.1136/bmj.h1793.

Sidman, M. (1960). *Tactics of Scientific Research: Evaluating Experimental Data in Psychology*. New York: Basic Books.

Smith, J.D. (2012). Single-Case Experimental Designs: A Systematic Review of Published Research and Current Standards. *Psychological Methods, 17*: 510-550.

Smith, J.D., Eichler, W.C., Norman, K.R., and Smith, S.R. (2014). The Effectiveness of Collaborative/Therapeutic Assessment for Psychotherapy Consultation: A Pragmatic Replicated Single-Case Study. *Journal of Personality Assessment*. Prepublication online at doi: 10.1080/00223891.2014.955917.

Smith, M.L., and Glass, G.V. (1977). Meta-Analysis of Psychotherapy Outcome Studies. *American Psychologist, 32*: 752-760.

Sterne, J.A.C. (Ed.). (2009). *Meta-analysis in Stata: An Updated Collection from the Stata Journal*. College Station, TX: Stata Press.

Swaminathan, H., Rogers, H.J., and Horner, R.H. (2014). An Effect Size Measure and Bayesian Analysis of Single-Case Designs. *Journal of School Psychology, 52*: 105-122.

Tasky, K.K., Rudrud, E.H., Schulze, K.A., and Rapp, J.T. (2008). Using Choice to Increase On-Task Behavior in Individuals With Traumatic Brain Injury. *Journal of Applied Behavior Analysis*, *41*(2): 261-265. doi: 10.1901/jaba.2008.41-261.

Tate, R., Perdices, M., Rosenkoetter, U., McDonald, S., Togher, L., Shadish, W.R., Horner, R., Kratochwill, T., Barlow, D., Kazdin, A., Sampson, M., Shamseer, L., and Vohra, S. (2014). *The Single-Case Reporting Guideline in Behavioural Interventions (SCRIBE): Explanation and Elaboration*. Manuscript in preparation.

Tate, R. L., Perdices, M., Rosenkoetter, U., Wakima, D., Godbee, K., Togher, L., and McDonald, S. (2013). Revision of a Method Quality Rating Scale for Single-Case Experimental Designs and N-of-1 Trials: The 15-Item Risk of Bias in N-of-1 Trials (RoBiNT) Scale. *Neuropsychological Rehabilitation*, *23*, 619-638.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse. (2014a). *Procedures and Standards Handbook* (Version 3.0). Washington, DC: What Works Clearinghouse.

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2014b). *Students with Learning Disabilities Intervention Report: Spelling Mastery*. Washington, DC: What Works Clearinghouse. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/intervention_reports/wwc_spelling_mastery_012814.pdf.

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2014c). *Students with Learning Disabilities Intervention Report: Repeated Reading*. Washington, DC: What Works Clearinghouse. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/intervention_reports/wwc_repeatedreading_051314.pdf.

Van den Noortgate, W., and Onghena, P. (2008). A Multilevel Meta-Analysis of Single-Subject Experimental Design Studies. *Evidence-Based Communication Assessment and Intervention, 2*: 142-151.

Vevea, J. L., and Hedges, L.V. (1995). A General Linear Model for Estimating Effect Size in the Presence of Publication Bias. *Psychometrika, 60*, 419-435.

Viechtbauer, W. (2010). Conducting Meta-Analyses in R With the Metafor Package. *Journal of Statistical Software, 36*(3): 1-48 (http://www.jstatsoft.org/v36/i03/)

Vohra, S., Shamseer, L., Sampson, M., Bukutu, C., Tate, R., Nikles, J., Kravitz, R., Guyatt, G., Altman, D.G., Moher, D., for the CENT group. (2014). *CONSORT Statement: an Extension for N-of-1 Trials (CENT). British Medical Journal.* doi: http://dx.doi.org/10.1136/bmj.h1738.

Wendt, O., and Miller, B. (2013). Quality Appraisal of Single-Subject Experimental Designs: An Overview and Comparison of Different Appraisal Tools. *Education and Treatment of Children, 35*: 235-268.

Wilkinson, L., and the Task Force on Statistical Inference. (1999). Statistical Methods in Psychology Journals: Guidelines and Explanations. *American Psychologist, 54*: 594-604.

Wilson, D.B. (2006). *Meta-analysis Macros for SAS, SPSS, and Stata*. Retrieved May 22, 2013, from. http://mason.gmu.edu/~dwilsonb/ma.html

Zelinsky, N.A.M., and Shadish, W.R. (in press). Choice-Making on Challenging Behaviors in People With Disabilities: A Meta-Analysis. *Developmental Neurorehabilitation*.

Zuur, A. F., Saveliev, A. A., and Ieno, E. N. (2012). *Zero Inflated Models and Generalized Linear Mixed Models With R*. Newburgh, UK: Highland Statistics.

Glossary

Accessibility: Refers to the ease with which the statistic can be used by researchers and practitioners who are generally not users of advanced statistical methods.

Alternating Treatments Design: A design in which different conditions (e.g., baseline and treatment) are alternated over time. Typically that alternation is more rapid than in the reversal design, for example, involving only one observation per phase rather than multiple observations within a phase.

Autocorrelation: The correlation of one observation in a time series with one or more observations that occurred previously.

Bayesian: Refers to a set of statistics tracing its history back to the Reverend Thomas Bayes, with a different set of rationales and statistical procedures than the statistics that are familiar to most researchers (the latter often called frequentist statistics, such as traditional regression). Bayesian statistics tend to have certain useful advantages and characteristics, for example, (1) they tend to function well in smaller sample sizes than frequentist statistics often require, (2) they allow the researcher to draw conclusions about the probability that a model is true in the population (contrary to common beliefs, frequentist statistics do not allow this), and (3) they allow (require) the researcher to specify prior knowledge about the parameter into the computations.

Between-Case Effect Size: A statistic describing the size of an effect using data from many cases, usually taking variability between cases into account (perhaps in addition to variation within-cases). See also Within-Case Effect Size.

Bias: A systematic deviation of an estimate from the population value. To be distinguished in particular from random error, which is a random deviation from the value.

CENT: An acronym for CONSORT Extension to N-of-1 Trials, these are guidelines to improve the reporting of N-of-1 trials.

Changing Criterion Design: A design in which levels of performance on the outcome that qualify for reinforcement are changed over time, typically made more stringent so that performance improves over time.

Confidence Interval: In frequentist statistics, an interval that surrounds an estimate of a parameter and that tells the probability that the parameter would fall into the interval when the study is repeated a large number of times. For instance, a 95% confidence interval suggests that if a study were repeated 100 times, 95 of the resulting 100 confidence intervals would contain the population parameter. Credible intervals in Bayesian statistics allow stating the probability (e.g., 95%) that a parameter lies in the interval.

Confounding: Occurs when the variable of interest systematically varies with an extraneous variable.

CONSORT: An acronym for Consolidated Standards of Reporting Trials, these are guidelines to improve the reporting of randomized controlled trials.

Counterfactual: In causal inference, the counterfactual is what would have happened to a unit that received a condition (usually a treatment) had that unit not received the treatment.

Covary: When two variables change systematically in relationship to each other, usually measured by a correlation or regression coefficient, or a covariance. In this example, treatment condition (e.g., whether a participant is in treatment or control) covaries with scores on the outcome variable—e.g., is being in treatment positively correlated with a good outcome.

Cumulative Meta-Analysis: A meta-analytic technique that shows how the average effect size changes over time as more studies are added to the meta-analysis.

$d_{between}$: A pedagogical term used in this paper to indicate any effect size (not just a *d*-statistic) standardized using data from more than one case.

d-statistic: Sometimes called the standardized mean difference statistic, computed as the difference in means of two conditions (e.g., the difference between the treatment group mean and the control group mean), divided by a measure of the standard deviation of that mean difference (often, the pooled standard deviation). Interpreted as a standard deviation, so that $d = 1$ would mean the treatment group did one standard deviation better than the control group on the outcome measure.

d_{within}: A pedagogical term used in this paper to indicate any effect size (not just a *d*-statistic) standardized using data from only one case.

Dependency: Most statistics assume independence of errors (or residuals). Some kinds of data are likely to violate that assumption, such as time series data in single-case designs where autocorrelation is likely.

Detrend: When trend exists in a time series, detrending is a statistical procedure to remove trend from the data prior to analysis. Many ways exist to detrend data.

Diagnostics: Statistical methods that shed light on whether the assumptions of a statistic are met or show how the statistic is influenced by particular observations in a set of observations.

Effect Size: An effect size describes the magnitude of an effect or relationship. An effect size estimator is a particular statistic for estimating an effect size parameter.

Evidence-Based Practice: The principle that practical decisions about what practices to use (e.g., in medicine, education, psychotherapy) should be based on the best available evidence from the research literature.

Fixed Effect: A parameter treated as fixed in the population, varying in the sample only by virtue of sampling error not by variation in the population.

Fixed Effects Model in Meta-Analysis: A model that assumes the researcher is interested only in inferences about the studies in the meta-analysis, in particular, inferences about how the results might vary if the sample of people were different.

Forest Plot: A graph that is widely used in meta-analysis, and that shows each study as a separate row. Typically, a dot represents a study effect size, the size of that dot reflects its weight in the meta-analysis, and the effect size is surrounded by a confidence interval. Many variations of these plots exist.

Functional Relationship: The conclusion that systematic covariation between outcome and conditions is due to a causal relationship between them.

Heterogeneity Testing: A set of statistics that examine the extent to which observed variation in effect sizes exceeds what would be expected by chance.

I^2 statistic: A descriptive statistic that shows the percentage of observed variance in effect sizes that exceeds what would be expected by chance.

Inferential Test: Tests that make inferences from samples to populations.

Internal Validity: The validity of an inference that an intervention caused an outcome.

Interrupted Time Series Design: An experimental design in which observations of the outcome are recorded sequentially over time, where an intervention is introduced at some point in the time series to see if the outcome changes.

Meta-Analysis: A set of quantitative methods for synthesizing the results of multiple studies on the same question.

Moderator: A variable that influences the size or direction of an effect or relationship.

Multilevel Model: A general term that refers to a class of statistics that can model dependencies among errors caused by autocorrelation, nesting, or other sources of dependency. Sometimes called hierarchical model or random coefficient model.

Multiple Baseline Design: A design reporting multiple time series, with each series starting treatment at a different point in time. Can be multiple cases, multiple outcomes on one case, or multiple settings for one case.

Multiple Effect Sizes within Study: Often a study will allow computation of more than one effect size. Such effect sizes cannot be presumed to be independent of each other and so risk violating the independence assumption if analyzed as if they are independent.

Nonequivalent Comparison Group Design: An experimental design in which the groups being compared (e.g., treatment and control) are not formed randomly.

Normality Assumption: The assumption that a variable (or the residuals of that variable around the mean) are normally distributed.

Overlap Statistics: In single-case designs, statistics that measure the extent to which observations from different phases within a case (e.g., baseline and treatment) overlap, where low overlap indicates higher treatment effectiveness. They are often called nonoverlap statistics to indicate that nonoverlap is associated with more effective treatments.

Percentage of Nonoverlapping Data: One of the earliest and most widely used of the overlap statistics, it measures how many observations from one condition (usually treatment) exceed the highest (or lowest, depending on the direction of a positive outcome) observation in another condition (usually baseline).

Phase: in a single-case design, a period of time in which the same condition is given to the case. Phases typically contain multiple observations in the most common SCDs (e.g., in reversal or multiple baseline design), but can have as few as one observation in alternating treatment designs.

Poisson Distribution: A distribution reflecting the probability of the occurrence of a given number of events, often used to represent count data.

Potential Outcomes Model: A model that posits that all units have multiple potential outcomes before an experiment begins, each potential outcome depending on which condition they would receive. Only after assignment occurs do we observe the actual outcome. The effect of a treatment is the difference between the two potential outcomes.

Power: The probability of correctly rejecting the null hypothesis. Usually described as the probability that the researcher will correctly conclude a treatment effect exists when, in fact, that effect does exist in the population.

Publication Bias: The hypothesis that the published literature is a biased sample of the complete set of studies ever done on a question and that relying only on published literature in a review may thus result in biased estimates of the answer to the question.

Q-Test Statistic: An inferential statistic to test for the presence of heterogeneity.

Quantile-Quantile Plot: A probability plot for comparing two distributions. Often used to see if a distribution of observed scores matches an underlying assumed distribution (e.g., normality).

R package: In the free computer program R (R Development Core Team, 2012), a package is a set of commands to perform certain functions or compute certain statistics. Packages are usually written by researchers who work on a specialized set of statistics not often available in the basic R program. They usually are not point-and-click, but rather require some use of syntax commands. They can be downloaded for free through the R website or associated sites.

Random Effect: A parameter hypothesized to vary in the population, allowed to vary by more than just sampling error.

Random Effects Model in Meta-Analysis: A model that assumes the researcher is interested in inferences that go beyond the studies in the meta-analysis, for example, what results might be if other variations in units, treatments, outcomes, or settings were used.

Randomized Experiment: An experiment in which conditions (e.g., treatment and control) are assigned by chance. In between-groups designs, conditions are usually assigned to units (e.g., cases), and in SCDs, conditions are usually assigned to time.

Regression Discontinuity Design: An experiment in which units are assigned to conditions based on a cutoff(s) on a measured variable prior to the experiment.

Reversal Design: A design in which multiple conditions (most frequently, a baseline condition and a treatment condition) are applied to a case in sequence, usually in phases with each phase having multiple observations (e.g., a baseline phase with five observations). The change from each phase to the next is a reversal.

SCRIBE: An acronym for Single-Case Reporting Guideline in Behavioural Interventions. These are guidelines to improve the reporting of single-case designs and N-of-1 trials for social, educational, and behavioral sciences.

Selective Reporting: When a report of research includes only some of the results that were computed, typically a biased rather than a random subset of results, as when only statistically significant results are reported.

Serial Dependence: In a time series, the fact that one observation may depend on the value of one or more previous observations.

Single-Case Design: An experimental design with repeated assessment of an outcome over time within a case, both in the presence and absence of an intervention, where the experimenter controls the timing of the intervention both within and across cases.

Small Sample Bias: The *d*-statistic overestimates the size of an effect in small samples, or more technically, in samples containing less statistical information; the bias can be corrected. This should not be confused with weighting by a function of sample size, which does not correct for small sample bias.

Source of Counterfactual Inference: Because the true counterfactual can never be observed, researchers rely on other sources of information to estimate the true counterfactual, such as a randomized control group, or an extrapolation of pretreatment observations.

Standard Error: Technically, the standard deviation of the sampling distribution of a statistic, used to measure how well a sample represents a population (small standard errors suggest the sample is likely to be closer to the population mean). Used to compute confidence intervals and statistical significance tests.

Standardized Effect Size: Any effect size index that puts study results on a scale that purports to have the same meaning across studies. One widely used class of standardized effect sizes are created by dividing a measure of effect by a standard deviation (as in the standardized mean

difference or the correlation coefficient). However our use of the term standardized includes a broader set of effect sizes that make results comparable across studies in different ways (such as risk ratios and odds ratios), including also the overlap statistics in SCDs.

Standardized Mean Difference: Often called a *d*-statistic (δ is its population parameter), this is an effect size obtained by subtracting the mean outcome of the comparison group from the mean outcome of the treatment group, and dividing that difference by an estimate of its standard deviation. The *d*-statistic is biased upwards (too large) in small sample sizes, and a correction for that results in the effect size statistic called *g*.

Statistical Significance: A decision that a statistic is unlikely to be observed by chance if the null hypothesis is true.

Tau-U: An overlap statistic that, unlike most such statistics, takes into account monotonically increasing or monotonically decreasing trend in the data.

Threats to Validity: A reason why an inference may be wrong; an alternative explanation for an observed effect or relationship.

Trend: A systematic increase or decrease in the value of observations over time. Trend can be linear or nonlinear.

Type I Error: An incorrect conclusion that an effect or relationship exists in the population when it does not. An example would be to conclude a treatment is effective when it is not. The most commonly used nominal Type I error rate is $\alpha = .05$.

Type II Error: An incorrect conclusion that an effect or relationship does not exist in the population when it does exist. An example would be to conclude a treatment is not effective when it is effective. The most commonly used nominal Type II error rate is $\beta = .20$.

Visual Analysis: A set of methods in SCD research to make inferences based on visual inspection of graphs of SCD data (e.g., inferences about the presence of trend or a treatment effect).

Wait-List Control: A control group in which participants are placed on a wait-list to receive the treatment after the experiment is finished.

Weighted and Unweighted Average Effect Size: When more than one effect size exists on a question, one can take the average of them in two ways. An unweighted average is the usual average (the sum of observations divided by the number of observations). A weighted average gives more weight to some samples than others. In meta-analysis, the most common weight is a function of sample size (usually inverse variance weighting) so that effects from studies with larger samples get more weight in the average.

Within-Case Effect Size: A statistic describing the magnitude of an effect for one case using data only from that one case, taking into account only variation in outcome within that one case. See also Between-Case Effect Size.

Zero-Inflation: When the outcome variable is a count, it can be modeled with a Poisson distribution. In some data sets, however, counts of zero are more prevalent than predicted by that distribution. In those cases, analyses that take this excess of zeroes into account are available (e.g., Zuur, Saveliev, and Ieno 2012).