EDUCATION AND TREATMENT OF CHILDREN Vol. 35, No. 2, 2012

Considerations for the Systematic Analysis and Use of Single-Case Research

Robert H. Horner
University of Oregon

Hariharan Swaminathan and George Sugai
University of Connecticut

Keith Smolkowski
Oregon Research Institute

Abstract

Single-case research designs provide a rigorous research methodology for documenting experimental control. If single-case methods are to gain wider application, however, a need exists to define more clearly (a) the logic of single-case designs, (b) the process and decision rules for visual analysis, and (c) an accepted process for integrating visual analysis and statistical analysis. Considerations for meeting these three needs are discussed.

Single-case research methods provide a scientifically rigorous approach for documenting experimental control (internal validity) and advancing effective practices in education, behavior analysis and psychology (Hersen & Barlow, 1976; Kazdin, 1982; Kennedy, 2005; Kratochwill, 1978; Kratochwill & Levin, 1992; McReynolds & Kearns, 1983; Richards, Taylor, Ramasamy & Richards, 1999; Tawney & Gast, 1984; Todman & Dugard, 2001). The long history of conceptual and clinical contributions from single-case research is encouraging (e.g., Dunlap & Kern, 1997; Odom & Strain, 2002; Sidman, 1960; Wolery & Dunlap, 2001) but underutilized as we consider the standards by which a new model for educational science is being defined (Kratochwill & Stoiber, 2000; Odom et al., 2005; Whitehurst, 2003). If findings from single-case research are to reach the larger scientific and

Correspondence to Robert H. Horner, Dept. of Special Education, College of Education, 5261 University of Oregon, Eugene, OR 97403-5261; e-mail: robh@uoregon.edu.

This paper was commissioned by the Institute of Education Sciences and supported in part by a grant from the Office of Special Education Programs, with additional funding from the Safe and Drug Free Schools Program, U.S. Department of Education (H326S030002). Opinions expressed herein are those of the authors and do not necessarily reflect the position of the U.S. Department of Education, and such endorsements should not be inferred. The authors extend appreciation to Dr. Cynthia Anderson, Dr. Charles Greenwood, Dr. Larry Hedges, and Dr. Thomas Kratochwill for comments on early drafts of this paper.

Pages 269-290

clinical community, however, a more transparent process is needed for (a) defining the logic guiding single-case methods, (b) describing the decision-rules guiding visual analysis of single-case data, (c) defining professional standards for using single-case research to identify empirically supported interventions, and (d) extending analysis of single-case results to include both visual and statistical methods.

We begin by revisiting the rationale for considering the addition of statistical analysis to the standard visual analysis protocols for single-case results. We place this discussion in the context of three emerging standards for assessing any research report. We then summarize decision-rules that may help standardize visual analysis. Finally, we extend the logic from these rules to considerations that might guide the development and use of formal effect-size indices.

A Rationale for Re-considering Statistical Analysis of Single-case Research

Statistical analysis of single-case data has been proposed to both define the statistical significance of observed effects, and document the effect size in a manner that facilitates meta-analyses. To date, however, no statistical approach for examining single-case research has met three fundamental criteria: (a) controls for auto-correlation (e.g. the fact that scores are not independent), (b) provides a metric that integrates the full constellation of variables used in visual analysis of single-case designs to assess the level of experimental control demonstrated by the data, and (c) produces an effect-size measure for the whole study (as opposed to two adjacent phases). Different proposals have succeeded in elements of this task (Busk & Serlin, 1992; Parker & Hagan-Burke, 2007; Parker, Hagan-Burke & Vannest, 2007), but none has offered a model that meets all three criteria. A full review of the history, considerations, and frustrations with the application of statistical analysis to single-case research is under development by two research projects funded by the U.S. Department of Education's Institute of Educational Sciences.

We argue that two emerging trends in the field make the need to reconsider statistical analysis of single-case data particularly important. The first is a new model for educational research that is based on high experimental rigor with emphasis on meta-analysis. With the founding of the Institute of Education Sciences (U.S. Department of Education, Education Science Reform Act, 2002), the U.S. Department of Education has launched a major initiative focused on guiding educational innovation based on scientific findings and ensuring that research in education meets the highest standards of science. One element of this effort is increased attention to the value and utility of meta-analyses. A meta-analysis allows the integration of a large body of experimental research and distillation of core themes for a field of study (Hedges, 2007; Hedges, Shymansky & Woodworth, 1989; Hunter & Schmidt, 2004). Meta-analyses are particularly valuable for policy makers attempting to define the applied implications from a body of research. A basic metric in meta-analysis is the size of experimental effects. Effect-size measures for traditional randomized control trials are well established, and multiple options exist (Cohen, 1988; Hedges, 2007; Lipsey & Wilson, 1993; Rosenthal, Rosnow, & Rubin, 2000; Rosnow & Rosenthal, 1996). Usable effect-size measures for single-case research are less well established and consensus on how to apply effect-size measures to single-case research has not yet emerged (Parker et al., 2007). As elaborated below, we believe that development of an effective strategy for incorporating single-case research findings into meta-analyses is essential for these findings to inform larger policy, research and dissemination efforts (especially outside the Behavior Analysis research community). As such, the need for agreement on an effect-size metric for single-case research is increasingly acute.

The second trend prompting the need for statistical analysis of single-case research is the growing focus on evidence-based practice and empirically-supported treatments (Logan, Hickman, Harris, & Heriza, 2008; Kratochwill & Stoiber, 2002; Odom et al., 2005; The Evidence-based Intervention Work Group, 2005). The basic thesis is that federal and state governments should invest with greater confidence in educational, clinical, and social practices that have been validated through rigorous research. Without inhibiting innovation and creativity, the commitment to empirically-supported treatments asserts that clinical interventions should be demonstrated to be feasible, acceptable and causally related to valued outcomes before society is asked to invest in widespread dissemination/implementation (Buvinger, Evans & Forness, 2007; Fixsen, Naoom, Blase, Friedman, & Wallace, 2005; Flay et al., 2005; Odom et al., 2005). An array of federal and professional councils are currently engaged in efforts to operationalize the criteria for empirically-supported treatments and to identify those treatments that should garner attention, resources and investment by society (e.g., Flay et al., 2005; What Works Clearinghouse, 2007).

A major challenge exists for defining the criteria by which single-case research may be used to identify empirically-supported treatments (Buvinger et al., 2007; Forness, Kavale, Blum & Lloyd, 1997; Horner et al., 2005). The addition of statistical models for analysis of single-case research, especially measurement of effect size, offers significant potential for increasing the use of single-case research in documentation of empirically-supported treatments (Parker et al., 2007; Van den Noortagate & Onghena, 2003).

Three Standards in Analysis of Research for Policy Impact

The interpretation of research results is both a science with documented tools, methods, procedures, and criteria and an art with many levels of perspective and application that are shaped by nuance and experience. The range of approaches to examining research results is part of what leads science to new and useful findings. Increasingly, however, experimental research in education, psychology and other social sciences is expected to respond to at least three core interpretive questions: (a) Is an experimental or causal effect documented? (b) What is the size or magnitude of the effect? and (c) Is the experimental effect socially important? Each of these questions is worthy of consideration from a single-case design perspective and as a basis for shaping the role of statistical models for summarizing and evaluating the outcomes of single-case research.

Is there documentation of an experimental or causal effect?

Experimental research – both group and single subject – typically defines one or more independent variables to be manipulated and one or more dependent variables to be monitored. A central focus of research design and analysis is documentation that change in the dependent variable is unlikely to have occurred for any reason other than the active (planned and controlled) manipulation of the independent variable (i.e., alternative explanations or hypotheses can be ruled out). An array of design and analysis procedures has been developed to allow this determination. The goal of these tools is to control for threats to internal validity (Campbell & Stanley, 1963; Shadish, Cook & Campbell, 2002). A strength of single-case designs and visual analysis is strong internal validity that allows documentation of experimental control through systematic and direct replication (Gast, 2010; Kazdin, 2011; Kennedy, 2005; Kratochwill, 1978).

What is the size of the experimental effect?

A more recent expectation of experimental research is that the research design and analysis not only demonstrates experimental control, but also quantifies the size of the demonstrated effect. Use of research results for policy purposes often focuses as much on the size of an experimental effect as on whether the effect is demonstrated through experimental control (Busk & Serlin, 1992; Hedges, 2007). Both group designs and single-case designs could provide powerful documentation of experimental control even if the size of the effect is modest (Hedges, 2007). Comprehensive interpretation of research findings includes determining not only the level of experimental control, but the size of the effect.

In either group-designs or single-case designs the goal is to assess the size of the experimental effect under analysis (Hedges, 2007). In group design research, quantification of size is documented through comparison of Treatment versus Control groups (Cohen, 1988; c.f. Parker & Hagan-Burke, 2007 for an expanded interpretation of effect size). For single-case research, assessment of the size of the treatment's effect includes not only estimating the effect size based on data from two adjacent phases (e.g. Baseline, Intervention), but more importantly the effect size of the functional relation based on data from the full set of phases used to document experimental control. We elaborate on this topic later in this paper.

What is the social importance of the finding?

The third question asked of any applied study is the classic query posed frequently in doctoral thesis orals, "so why should we care about this finding?" The social validity, or social importance, of research findings will vary depending on many features (e.g. the magnitude of the effect, the value of the effect for social outcomes, the feasibility of achieving the effect in typical social contexts). While economists, ethicists, scientists and decision-theorists (Gresham & Lopez, 1996; Witt & Martens, 1983; Wolf, 1978) offer guidelines for addressing this third question, the reality is that the value of research findings will often rest on variables outside the formal research design (e.g., comparison with local or contextualized social norms). As such, we do not focus our attention on this question in this paper. However, see Strain, Barton, and Dunlap in this issue.

Using Visual Analysis to Document Experimental Control in Single-case Research

The process by which single-case designs document internal validity (experimental control) is well established in an array of texts (Gast, 2010; Hersen & Barlow, 1976; Kazdin, 1982, 2011; Kennedy, 2005; Kratochwill & Levin, 1992; Sidman, 1960; Tawney & Gast, 1984). Our goal here is not to review the broad structure and logic of single-case methods, but to focus on the specific features of visual analysis that may warrant replication/consideration in the development of any comprehensive statistical model. A need exists to be more explicit about the decision-rules used to conduct visual analysis of single-case data, and use the results from that analysis to infer a functional relation. Emphasis is placed on the two most commonly used single-case designs (i.e., reversal/withdrawal: A [Baseline 1] B [Intervention 1] A [Baseline 2] B [Intervention 2] and multiple baseline), because they exemplify the core features of single-case methods, especially repeated replications of effects over time and across conditions. Case studies and simple interrupted time series designs such as AB or BA studies are not considered experimental designs and are not included in this discussion. Other single-case designs (e.g. alternating treatments/ multi-element; changing criterion) will be addressed by extension in future papers. For purposes of the present discussion, we propose that studies using ABA reversal and two-series multiple baseline designs be viewed as acceptable for documenting experimental control only with reservation and under specific conditions. Though these designs have been traditionally viewed as minimally acceptable for demonstrating experimental control (Kennedy, 2005), we suggest that a more rigorous standard may benefit the field.

The fundamental logic behind single-case experimental research is documentation that active, planned, and systematic manipulation of an independent variable is functionally related to predicted and observed change in the dependent variable. To achieve this outcome (a) each participant serves as his or her own control (the behavior of the individual participant is the unit of analysis), (b) the dependent variable is measured repeatedly across time and this measurement begins in a baseline phase before intervention to demonstrate a clear pre-intervention pattern of performance, and (c) the independent variable must be actively manipulated. Experimental control is demonstrated when manipulation of the independent variable is followed by a predicted change in the dependent variable at a minimum of three different points in time. Figure 1 provides examples of an ABAB reversal design and a four-series multiple baseline across participants design. In each design the circled numbers indicate how the design provides the opportunity to document (a) at least three demonstrations of change in the dependent variable following active manipulation of the independent variable (four demonstrations in the multiple baseline design), (b) across at least three different points in time. This standard for demonstrating a functional relation is conservative, yet feasible. It is a standard we believe will have a high likelihood of acceptance by the larger academy of researchers, policy-makers, and clinicians concerned with human behavior.

As with group research, it is the *design* of single-case research (i.e., arrangement of conditions and manipulation of the independent variable), not the mode of analysis (i.e., visual or statistical), that is critical for documentation of experimental control (Bloom, Fischer & Orme, 1999). For single-case research the standard for documenting experimental control, three demonstrations of predicted effect at three different points in time, is based on emerging professional convention (Browder, Spooner, Ahlgrim-Delzell, Harris & Wakeman, 2008;

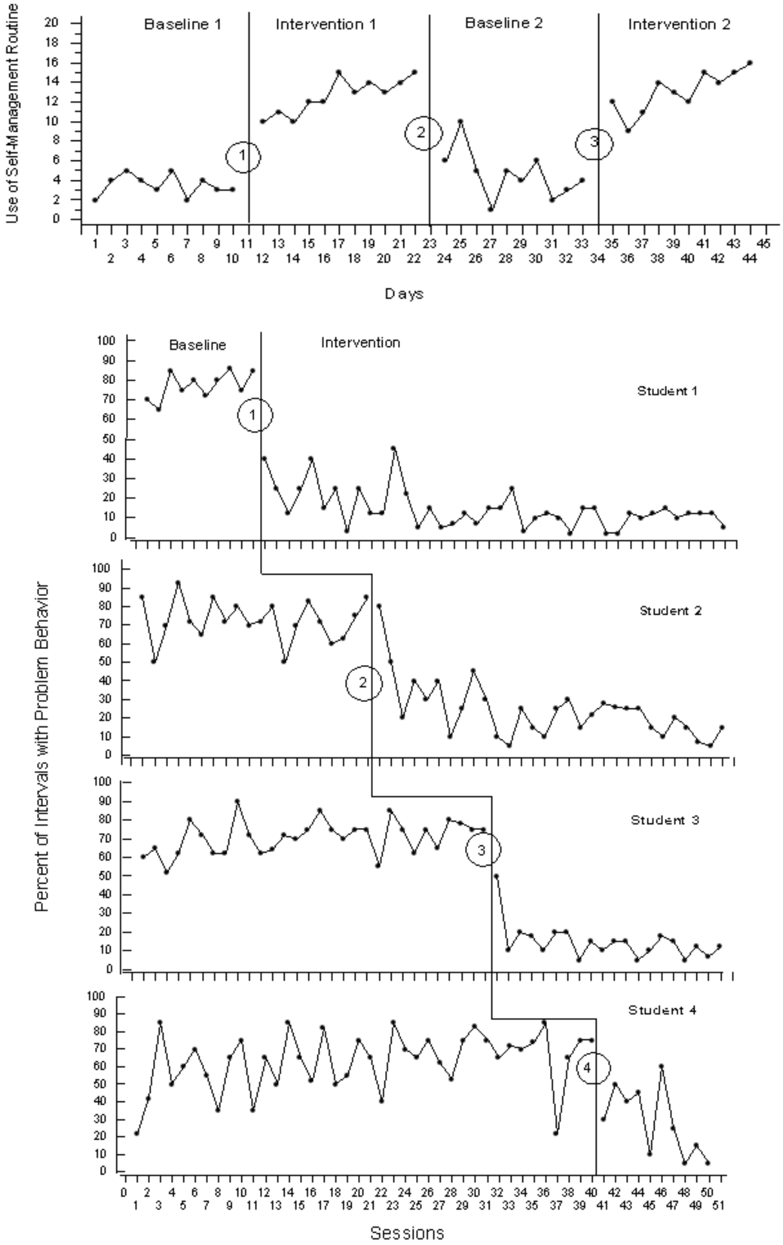

Figure 1: ABAB and Multiple Baseline Design documenting at least 3 demonstrations of effect (3 for ABAB; 4 for Multiple Baseline).

Horner et al., 2005; Kratochwill et al., 2010; Kratochwill, Levin, Horner & Swaboda, 2011; Lane, Kalberg & Shepcaro, 2009; Roane, Rihgdahl, Kelley & Glover, 2011). ABA and two-series multiple baseline designs allow two demonstrations of an effect each at a different point in time. While two, independent manipulations has been viewed as minimally adequate to demonstrate experimental control in the past, we propose that the more stringent, "three-demonstration" standard offers more confidence in documenting an experimental effect. All other things being equal, an increase in the number of demonstrations of an effect increases our confidence in the degree of experimental control.

The process of assessing experimental control in single-case designs typically begins with comparison of data from adjacent phases and continues with integration of data patterns from all phases in the study. To assess if an effect is demonstrated across any two phases (e.g. determine if introduction, or removal, of an intervention produced a predicted change in the dependent variable), data from the second phase are compared initially with the data from the first phase and then with the "projected results" (e.g., extension of the data pattern from the first phase into the second phase). Examples of adjacent phases from two studies with differing data patterns are provided in Figure 2. In each case, data in the second phase are examined and compared (a) with the actual data from the first phase and (b) with the expected, or projected, data pattern (with confidence intervals) obtained by extending data from the first phase into the second phase (the shaded areas in Figure 2). Visual analysis of data involves the simultaneous assessment of the level, trend, and variability of the data within and across adjacent phases. When data from two adjacent phases are compared, the rules of visual analysis also include assessment of immediacy of effect, the level of overlap, and the consistency of data patterns in similar phases (Parsonson & Baer, 1978). The role of each of these variables in visual analysis is outlined below.

Baseline

A central feature of single-case research is active documentation of performance under "treatment as usual" or Baseline conditions. Most single-case designs begin with a Baseline phase that includes at least five data points and establishes (a) the current pattern of responding and (b) a confident prediction of the pattern of future responding. Detailed description of the context in which Baseline data are collected (e.g., physical location, personnel, curricula, contingencies) is important, because the single-case experimental design focuses on the effects of manipulating the independent variable while all other Baseline variables are held constant. Confidence that uncontrolled

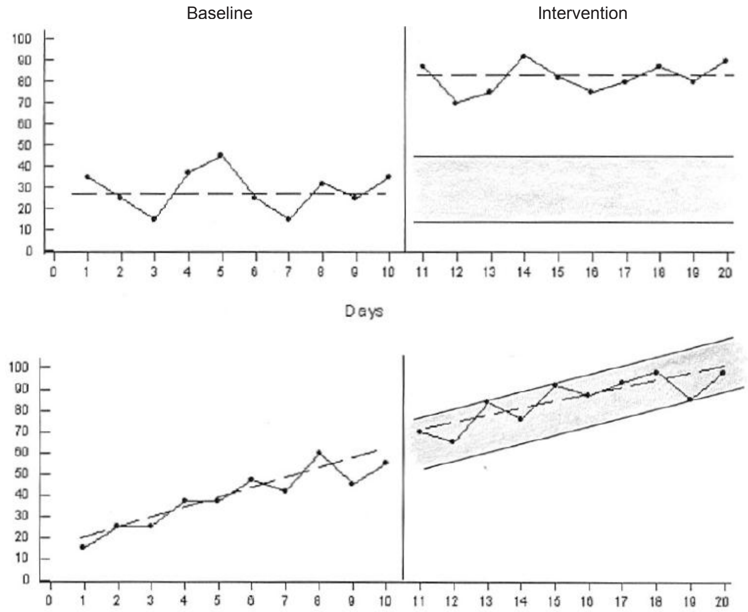

Figure 2: Baseline and Interventions phases from a single-case design. Shaded area indicates the “expected” or “projected” data (with confidence intervals) in the Intervention phase based on data in the Baseline phase.

variables are unlikely to be responsible for change in the dependent variable requires both careful description of the Baseline context and assurance that only the independent variable was altered at the point of intervention.

Change in level

The level of a phase typically refers to a measure of central tendency (e.g., mean, median) of all data points within the phase. An initial assessment of effect begins by comparing the level of the data in the first phase to the level of data in the second phase. This process is repeated by comparing the level of the observed data in the second phase with the level of the projected data from the first phase. Level change in visual analysis is affected by the degree of within-phase variability, within-phase trend (slope), data overlap, and change in trend between phases. As a general rule, more data points are required to demonstrate an effect where there is more within-phase variability, within-phase trend, or between-phase overlap.

Change in trend

The trend of data within a phase is represented by the slope of the line that best fits the data (i.e., either linear or quadratic). The data within a phase are considered more stable when the data points are closer to the trend line. The greater the difference in the slope of trend lines between adjacent phases, the greater confidence we have in the differences in responding between those phases. For example, a slightly increasing trend in phase 1 followed by a flat trend in phase 2 would be viewed as documenting a weaker effect than if the trend in phase 2 was slightly decreasing. In general, the weight given to the importance of trend within and between phases is affected by trend direction, within phase stability, and number of contributing data points. As the stability of a within-phase trend line becomes more difficult to interpret (e.g., extended phase lengths, curvilinear response pattern within a phase), more interpretive weight is given to the last three-to-five data points. Data sets based on "steady states" (low trend) require fewer data points to document a predictable pattern than do "transition states" (higher trend) (Sidman, 1960).

In addition to examining changes in trend based on the actual obtained data, trend in one phase can also be used to project expected data from phase 1 into phase 2 (see Fgure 2). If, for example, the trend line in phase 2 approximates the extended or predicted trend line from phase 1, a "no differences" interpretation would result (e.g., bottom panel in figure 2). However, if the slope of the projected trend line was reversed, or if there was a noticeable level change between the projection based on phase 1 and the actual data in phase 2, then statements about differences between phases 1 and 2 can be made with greater confidence. As noted below, this analysis is important to consider when attempting to establish an effect-size value for the overall study.

Change in variability

Variability of data within a phase refers to deviation in scores about the trend line. In the past, the range of scores within a phase has been used as an index of variability, but this approach is limited to phases that have minimal trend. Variability from the trend line is a more general indicator of variability – and is relevant whatever the slope of the trend line. The greater the variability within a phase, the larger the number of data points needed to document a predictable within-phase pattern. Change in variability between phases is a potential indicator of a treatment effect, even if no changes in level and trend are observed.

As with level and trend, variability is assessed both in terms of observed data in phase 2 compared to observed data in phase 1, and observed data in phase 2 compared to data *projected* from phase 1 into phase 2. Variability of data within a phase need not be uniform throughout the phase. For example, variability might be high initially and then become smaller across the observations within a phase. In this case, the variability in the latter part of the phase would be weighted more highly in visual analysis, and the projected variability would be much smaller than the overall variability observed in phase 1.

Degree of overlap

Overlap refers to the proportion of data points in phase 2 that overlap with the data from phase 1. This kind of comparison has been used recently in assessments of effect size (Percent of Non-overlapping Data, PND, Busk & Serlin, 1992), but it has held a traditional role in the basic assessment of an effect between two phases. Low overlap suggests a larger effect, and the interpretive value or weight of overlap in visual analysis is greatest when trend and variability are minimal. High within-phase variability or significant trends tend to reduce the weight given to overlap as a measure of effect. For example in the bottom panel of figure 2, the two phases have increasing trend with no overlap, yet the data in phase 2 are clearly predicted by projecting the trend from phase 1. In this case, the absence of overlap, typically a pattern associated with an effect, would not indicate that the introduction of the independent variable was associated with change in the pattern of the dependent variable.

Immediacy of effect

A final variable in the visual analysis of an effect between two phases is the immediacy of any change in data patterns following manipulation of the independent variable. Immediacy of effect may be calculated as the mean or median difference between the last three-to-five data points from phase 1 and the first three-to-five data points in phase 2. In many instances, researchers would predict gradual rather than rapid shift in the dependent variable (e.g., development of a complex motor skill, acquisition of a multiple element academic concept), and in these cases "immediacy of effect" is less weighted in the analysis.

In general, the greater the immediacy of effect the more likely the change is associated with manipulation of the independent variable. The immediacy of effect is assessed both with the observed data from phases 1 and 2, and a comparison of the observed data in phase 2 with the projected data from phase 1.

Defining experimental control based on results from all phases

A central principle in visual analysis is that while comparisons of adjacent phases are essential for interpreting a study, the assertion of experimental control only results from all data from all phases examined collectively. Taken together, the phases of a single-case study need to document three demonstrations of an effect at three different points in time. The consistency, replication, and magnitude of the effect determine the level of confidence with which experimental control is claimed.

In visual analysis of single-case designs, as with statistical analysis of group designs, the documentation of experimental control often is described as a dichotomous variable (present or absent) while in practice experimental control lies on a more continuous scale. Experimental control can be established with higher and lower levels of confidence. A major strength of single-case research, however, is the de-emphasis on data summary, transformation or reduction and the emphasis on descriptive display of actual raw data. This gives the reader all the relevant data and tools to independent judge the level of experimental control.

Vertical analysis in multiple baselines

Vertical comparison of data in each series of a multiple baseline is a less frequently mentioned criterion for assessing experimental effects. Vertical analysis is the assessment of whether changes in the DV for one series following IV manipulation (e.g. Participant #1) are associated with *no changes* in the other series where the IV is NOT manipulated (e.g. Participants #2, #3 and #4). In Figure 3, the lower design from Figure 1 is replicated to illustrate a multiple baseline design with four participants. This design allows four demonstrations of an effect at four points in time across series (i.e., participants). Visual analysis of this design would begin with phase by phase comparisons (Baseline to Intervention) within each series to assess the effect for each participant. As described above, comparison of baseline data to intervention data would occur through review of level, trend, variability, overlap and immediacy of effect for both observed and projected data.

Visual analysis of this multiple baseline, however, would include an additional "vertical" assessment to determine internal validity. When the intervention is implemented in the first series (e.g., with participant #1), visual analysis includes assessing not only if change (level, trend, variability, overlap) is observed in the subsequent phase within the series (participant #1), but also if *absence of change* is seen in the non-intervened series (participants 2, 3, and 4) where the independent variable has not been manipulated (see the vertical box in

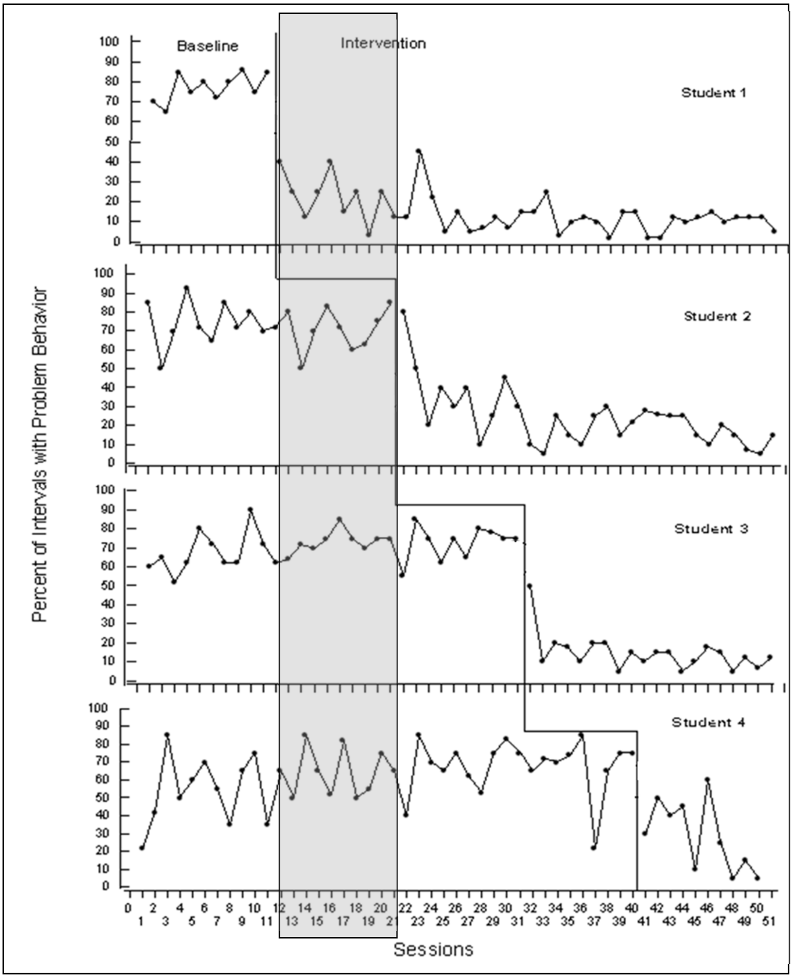

Figure 3: The box within this multiple baseline design indicates an example of the vertical analysis used to determine if (a) change in the dependent variable following intervention in the one series (session 13-21) is associated with (b) no change in the date for the other three series.

Figure 3). If change occurs in one series and not in others, support is increased for the inference that the change was due to manipulation of the independent variable rather than to uncontrolled events. Vertical analysis is used to assess the effect at each intervention point in the multiple baseline design. In Figure 3 the vertical analysis would be replicated for each participant; the data pattern immediately after the student experienced the intervention would be compared to simultaneous data from all the other participants. The claim for experimental control is supported if the data change *only* when the intervention has been applied.

Limitations of visual analysis

Taken together the multiple elements of visual analysis can be used to reliably index the extent to which data from a full single-case design document experimental control (Kennedy, 2005; Parsonson & Baer, 1978). The frequently cited demonstrations of poor agreement in visual analysis typically have focused on comparisons of two phases, and seldom have asked the relevant questions about consistency in visual analysis when all phases of a study are included (Brossart, Parker, Olson, & Mahadevan, 2005; Harbst, Ottenbacher, & Harris, 1991; Ottenbacher, 1993). The continued use of visual analysis to interpret single-case designs lies not in unwillingness of single-case researchers to respond to demonstration of poor inter-observer agreement but (a) the small number of inter-observer agreement studies that use the full data sets from a study for interpretation by trained researchers, (b) on-going publication of research applying visual analysis tools, and (c) the absence of accepted statistical alternatives.

Even if the utility of visual analysis procedures to determine functional relations is accepted, a challenge remains to define how single-case research can best contribute to the larger knowledge-base in education. Single-case research is especially well suited, for example, to development and refinement of new interventions prior to the investment in group design comparisons. Single-case research designs also fit well when interventions target low-incidence populations, especially when variability is high in the core measures for individuals with low-incidence labels (e.g., autism spectrum disorder). A challenge remains, however, when attempts are made to combine the findings from several single-case studies, or combine the knowledge gleaned from single-case research with findings from group-design research. To meet this challenge a functional index of effect size is needed for single-case results.

The Need for Single-case Research Effect-Size Measure

For single-case research results to be included in formal meta-analyses, and contribute to efforts to define empirically-supported treatments, some reliable index of effect size is needed (Busk & Serlin, 1992; Hedges, 2007; Parker & Hagen-Burke, 2007; Parker et al., 2007). The established protocol for identifying empirically-supported treatments has been to assess the average effect size of the treatments across a body of group design studies (Hedges, 2007). Single-case studies seldom are evaluated in terms of "effect size" and no common standard for meta-analysis has been accepted. The reasons we need a measure of effect size lie in (a) emphasizing the importance of not only demonstrating experimental control, but documenting also the *size of the effect*, and (b) in facilitating the meta-analysis process in a manner that will be compelling beyond the current behavior analytic community. The growing array of recent efforts to define effect-size indices for single-case research is a response to this need (Huitema & McKean, 2000; Parker et al., 2005; Parker et al., 2007). The intuitively accessible Percent of Non-overlapping Data (PND) was among the first effect-size measures proposed (Scruggs, Mastropieri, & Casto, 1987) and is the most frequently applied index at this time (Parker & Hagan-Burke, 2007). Parker et al. (2007) recently offered a variation on this theme, the Percentage of All Non-Overlapping Data (PAND) which has the advantage of allowing translation into a more interpretable form as Pearson's *Phi* or *Phi*^{2}. Busk and Serlin (1992) proposed use of a Standard Mean Difference (SMD) between phases to assess magnitude of effects, and Allison and Gorman (1993) recommend use of R^2 to assess effect size in single-case designs. As the need has grown for including single-case research in formal meta-analyses, interest has also increased in application of hierarchical linear modeling approaches to calculating effect size with single-case research (Marquis et al., 2000; Van Den Noortgate, & Onghena, 2003; Shadish, Rindskopf & Hedges, 2008).

The importance of these efforts warrants an entire paper focused only on the current state of the field and the strengths and limitations of each proposed option (Shadish et al., 2008). However, as statistical models are designed to fit the logic and structure of single-case research, we emphasize the importance of developing and applying effect-size measures that have the following features:

a) The effect-size index is comparable with Cohen's *d* making it easily interpreted and readily integrated into meta-analyses. We anticipate that the value of single-case effect-size measures will be most important for meta-analyses that include group and single-case studies or with clusters of single-case studies. The application of Cohen's *d* as a common standard in group design research, makes it useful to identify a comparable metric that would fit single-case methods.

b) The effect-size index reflects the experimental effect under analysis (e.g., all phases of a single-case study used to document experimental control) and not simply analysis of two adjacent phases. The conceptual logic of effect-size measurement for single-case methods needs to remain focused on assessing the magnitude of the functional relation. Given that a critical principle in single-case analysis is use of all data from all phases in a study to determine the functional relation, any measure of effect size needs to similarly include consideration of all data across all phases of the study.

c) The effect-size index controls for (1) serial dependency (a challenge with parametric analyses), (2) score dependency (a challenge for analyses using Chi Square models), and (3) variation in the weight given to changes in level, trend, variability and overlap. These are high standards but important for avoiding previously failed efforts. Any effect-size measure needs to control for the fact that scores in a single-case study are not independent. The fact that repeated measures come from the same participant requires care in any assumption of score independence, and assumptions about score distribution. Similarly, assessment of effect size should reflect the unique and combined results from changes in the level, trend, variability and overlap of data patterns across phases.

We believe the need remains to identify an option for effect-size measurement that meets these criteria.

Implications for Documentation of Empirically-Supported Treatments in Education, Behavior Analysis and Psychology

A major goal in education today is documentation of educational treatments that have an unequivocal and causal relationship with valued educational outcomes. These empirically-supported treatments are identified through published research in peer-reviewed journals. While agreement is strong that randomized control trials provide the needed rigor to define empirically-supported treatments (Shavelson & Towne, 2002; Whitehurst, 2003), less consensus exists on the standards by which single-case research can be used to demonstrate empirically-supported treatments (Horner et al., 2005; Kratochwill & Stoiber, 2002).

Any logic for defining empirically-supported treatments will be somewhat subjective, yet value exists in articulating specific and conservative criteria. Building on prior recommendations (Horner et al., 2005; Kratochwill & Stoiber, 2002; Logan et al., 2008), we endorse the following as a set of standards for establishing empirically-supported treatments via single-case research:

1. The intervention or practice is operationally defined;
2. The intervention or practice is implemented in typical contexts by typical intervention agents;

3. The intervention or practice is implemented with documented fidelity;
4. The intervention or practice is examined within an experimental design that demonstrates a causal relation between the intervention (practice) and the targeted educational outcome(s);
5. The intervention or practice is assessed with experimental effect across a sufficient range of studies, researchers, and participants to allow confidence in the effect. We propose that this standard be operationalized as (a) a minimum of five peer-reviewed studies, (b) conducted across at least three different locations/research groups, with (c) at least 20 different participants.
6. If a useful index of effect size can be developed for single-case research we would add a final criterion that at least five studies are conducted with documented effect size equaling or exceeding .50.

Summary

As the field of education embraces the evidence-based practice movement, careful attention will be given to the standards by which any research approach documents a treatment as empirically-supported. Single-case research has established a long and impressive history of contributions to special and general education, yet the contributions from single-case research are not widely recognized. The thesis of this paper is that while visual analysis procedures remain appropriate for documentation of experimental control, formal systems for assessing effect size in single-case research will be needed for this knowledge to be more widely accepted.

References

Allison, D., & Gorman, B. (1993). Calculating effect sizes for meta-analysis: The case of the single-case. *Behavior Research and Therapy, 31,* 621-641.

Bloom, M., Fischer, J., & Orme, J. (1999). *Evaluating practice: Guidelines for accountable professionals (3rd ed.).* Needham Heights, MA: Allyn and Bacon.

Brossart, D., Parker, R., Olson, E., & Mahadevan, L. (2005). The relationship between visual analysis and five statistical analysis in a simple AB single-case research design. *Behavior Modification, 30,* 531-563.

Browder, D., Spooner, F., Ahlgrim-Delzell, L., Harris, A., & Wakeman S. (2008). A meta-analysis on teaching mathematics to students with significant cognitive disabilities. *Exceptional Children, 74*, 407-432.

Busk, P., & Serlin, R. (1992). Meta-analysis for single-participant research. In T. R. Kratochwill & J. R. Levin (Eds.). *Single-case research design and analysis: New directions for psychology and education* (pp.187-212). Mahwah, NJ: Erlbaum.

Buvinger, E., Evans, S., & Forness, S. (2007). Issues in evidence-based practice in special education for children with emotional or behavioral disorders. In S. Evans, M. Weist, & Z. Serpell (Eds.), *Advances in school-based mental health interventions: Best practices and program models* (Vol. II, pp.19-1 – 19-19). Kingston, NJ: Civic Research Institute.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago, IL: Rand McNally.

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

Dunlap, G., & Kern, L. (1997). The relevance of behavior analysis to special education. In J.L. Paul et al. (Eds.), *Foundations of special education: Basic knowledge informing research and practice in special education* (pp. 279-290). Pacific Grove, CA: Brooks/Cole.

The Evidence-based Intervention Work Group. (2005). *Psychology in the Schools, 42*, 475-494.

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation Research: A Synthesis of the Literature.* Tampa, FL: University of South Florida, Louis de la Parte Florida Mental Health Institute, The National Implementation Research Network (FMHI Publication #231), 11, 247-266.

Flay, B., Biglan, A., Boruch, R., Castro, F., Gottfredson, D., Kellam, S., Moscicki, E., Schinke, S., & Valentien, J. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention Science, 6,* 151-175.

Forness, S., Kavale, K., Blum, I., & Lloyd, J. W. (1997). Mega-analysis of meta-analysis: What works in special education and related services. *Teaching Exceptional Children, 29*, 4-9.

Gast, D. (2010). *Single subject research methodology in behavioral sciences*. New York, NY: Routledge.

Gresham, F., & Lopez, M. (1996). Social validation: A unifying concept for school-based evaluation, research and practice. *School Psychology Quarterly, 11,* 204-227.

Harbst, K. B., Ottenbacker, K. J., & Harris, S. (1991). Interrater reliability of therapists' judgments of graphed data. *Physical Therapy, 71,* 107-115.

Hedges, L. V. (2007). Meta-analysis. In C. R. Rao & S. Sinharay (Eds.), *The handbook of statistics.* Amsterdam, Netherlands: Elsevier.

Hedges, L. V., Shymansky, J. A., & Woodworth, G. (1989). *A practical guide to modern methods of meta-analysis*. Washington, D.C.: National Science Teachers Association.

Hersen, M., & Barlow, D. H. (1976). *Single-case experimental designs: Strategies for studying behavior change*. New York, NY: Pergamon.

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

Huitema, B., & McKean, J. (2000). Design specification issues in time-series intervention models. *Educational and Psychological Measurement, 60,* 38-58.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings (2nd Ed.).* Newbury Park, CA: Sage.

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings*. New York, NY: Oxford University Press.

Kazdin, A.E. (2011). *Single-case research design: Methods for clinical and applied settings.* New York, NY: Oxford University Press.

Kennedy, C. H. (2005). *Single-case designs for educational research*. Boston, MA: Allyn and Bacon.

Kratochwill, T. (Ed). (1978). *Single subject research.* New York, NY: Academic Press.

Kratochwill, T., & Levin, J. R. (1992). *Single-case research design and analysis: New directions for psychology and education*. Hillsdale, NJ: Lawrence Erlbaum.

Kratochwill, T., & Stoiber, K. (2000). Empirically supported interventions and school psychology: Conceptual and practical issues: Part II. *School Psychology Quarterly, 15,* 233–253.

Kratochwill, T., & Stoiber, K. (2002). Evidence-based interventions in school psychology: Conceptual foundations for the Procedural and coding manual of Division 16 and Society for the study of school Psychology Task Force. *School Psychology Quarterly, 17,* 341-389.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). Single case designs technical documentation. In What works Clearinghouse: Procedures and standards handbook (Version 2.0). Retrieved from What Works Clearinghouse website: http://ies.ed.gov.ncee.wwc.pdf.wwc_procedures_v2_standards_handbook.pdf.

Kratochwill, T. R., Levin, J. R., Horner, R. H., & Swoboda, C. M. (2011). Visual analysis of single-case intervention research: Conceptual and methodological considerations (WCER Working Paper No. 2011-6). Retrieved from University of Wisconsin-Madison, Wisconsin Center for Education Research website: http://www.wcer.wisc.edu/publications/workingpapers/papers.php.

Lane, K., Kalberg, J., & Shepcaro, J. (2009). An examination of the evidence-base for function-based interventions for students with emotional and/or behavioral disorders attending middle and high schools. *Exceptional Children, 75*, 263-281.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*, 1181-1209.

Logan, L., Hickman, R., Harris, S., & Heriza C. (2008). Single-subject research design: Recommendations for levels of evidence and quality ratings. *Developmental Medicine & Child Neurology*, *50,* 99-103.

Marquis, J., Horner, R., Carr, E., Turnbull, A., Thompson, M., Behrens, G.,... Doolabh, A. (2000). Meta-analysis of positive behavior support. In R. Gersten, E. Schiller & S. Vaughn (Eds.), *Contemporary special education research: Syntheses of the knowledge base on critical instructional issues (pp.137-178).* Mahwah, New Jersey: Lawrence Erlbaum Associates.

McReynolds, L., & Kearns, K. (1983). *Single-subject experimental designs in communicative disorders*. Baltimore, MD: University Park Press.

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

Odom, S., & Strain, P. S. (2002). Evidence-based practice in early intervention/early childhood special education: Single-subject design research. *Journal of Early Intervention*, *25*, 151-160.

Ottenbacher, K. (1993). Interrater agreement of visual analysis in single-subject decisions: Quantitative review and analysis. *American Journal on Mental Retardation, 98*, 135-142.

Parker, R., Brossart, D., Callicott, K., Long, J., Garcia De-Alba, R., Baugh, F., & Sullivan J. (2005). Effect size in single-case research: How large is large? *School Psychology Review 34*, 116-132.

Parker, R., & Hagan-Burke, S. (2007). Useful effect size interpretations for single-case research. *Behavior Therapy, 38*, 95-105.

Parker, R., Hagan-Burke, S., & Vannest, K. (2007). Percent of all non-overlapping data (PAND): An alternative to PND. *The Journal of Special Education, 40*, 194-204.

Parsonson, B., & Baer, D. (1978). The analysis and presentation of graphic data. In T. Kratchowill (Ed.), *Single subject research* (pp.101-166). New York: Academic Press.

Richards, S. B., Taylor, R., Ramasamy, R., & Richards, R. Y. (1999). *Single subject research: Applications in educational and clinical settings*. Belmont, CA: Wadsworth.

Roane, H., Rihgdahl, J., Kelley, M., & Glover, A. (2011). Single-case experimental designs. In W. Fisher, C Piazza & H. Roane (Eds.), *Handbook of applied behavior analysis (pp. 132-147).* New York, NY: The Guilford Press.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrast and effect sizes in behavioral research: A correlational approach*. New York, NY: Cambridge University Press.

Rosnow, R. L., & Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Psychological Methods, 1,* 331-340.

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

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton-Mifflin.

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

Shavelson, R., & Towne, L. (2002). *Scientific research in education*. Washington, D.C.: National Academy Press.

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology*. New York, NY: Basic Books.

Tawney, J. W., & Gast, D. L. (1984). *Single subject research in special education*. Columbus, OH: Merrill.

Todman, J., & Dugard, P. (2001). *Single-case and small-n experimental designs: A practical guide to randomization tests.* Mahwah, NJ: Lawrence Erlbaum Associates.

U.S. Department of Education. (2002). Education Sciences Reform Act. Washington, DC: Author.

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

What Works Clearinghouse.(2007). Retrieved from http://www.w.w.c.org.

Whitehurst, G. J. (2003). Evidence-based education. Retrieved from http://www.ed.gov/offices/OERI/presentations/evidencebase.ppt

Witt, J., & Martens, B. (1983). Assessing the acceptability of behavioral interventions used in classrooms. *Psychology in the Schools, 20*, 510-517.

Wolery, M., & Dunlap, G. (2001). Reporting on studies using single-subject experimental methods. *Journal of Early Intervention, 24*, 85-89.

Wolf, M. M. (1978). Social validity: A case for subjective measurement or how applied behavior analysis is finding its heart. *Journal of Applied Behavior Analysis, 11*, 203-214.