Contemp School Psychol (2015) 19:136–144
DOI 10.1007/s40688-014-0025-7

TOOLS FOR PRACTICE

Methods for Assessing Single-Case School-Based Intervention Outcomes

R. T. Busse · Ryan J. McGill · Kelly S. Kennedy

Published online: 17 July 2014

Abstract The purpose of this article is to present various single-case outcome assessment methods for evaluating school-based intervention effectiveness. We present several outcome methods, including goal attainment scaling, visual analysis, trend analysis, percentage of non-overlapping data, single-case mean difference effect size, reliable change index, and convergent evidence scaling. The strengths and limitations of each of these methods are discussed, along with the potential practice applications. We conclude the article with a suggested display format for single-case intervention outcomes.

Keywords Single-case design · Evaluation · Intervention · Outcomes · Effect size

Methods for Assessing Single-Case School-Based Intervention Outcomes

In the wake of changes to IDEA that allow for response to intervention models, and the current resurgence in the field of school psychology toward a problem-solving approach of service delivery, there is a need to advance our methods for assessing and evaluating intervention outcomes. Whereas our profession prides itself on our assessment knowledge and skills, these skills often are adequately applied only to problem identification or classification purposes. Treatment outcome assessment, on the other hand, often is approached in an unsystematic manner or is based on a single outcome indicator rather than on outcome assessment that includes multiple methods. Thus, what is deemed best practices in assessment may be overlooked when we engage in treatment evaluation, thereby limiting our data-based decision making processes and accountability.

The purpose of this article is to provide an overview of the methods for assessing and monitoring treatment outcomes that can be used in research and practice, with particular emphases on practical applications within school or clinical settings. These assessment methods include: goal attainment scaling, visual analysis, trend analysis, percentage of non-overlapping data, single-case mean difference effect size, reliable change index, and convergent evidence scaling. We describe each of these methods, along with the strengths and limitations of each approach. We conclude the article with case illustrations of a suggested format for displaying single-case outcomes that employs multiple outcome indicators.

Before we delve into the disparate outcome indicators, a brief conceptual perspective is warranted. In the realm of treatment outcomes, much of the research is based upon a statistical premise wherein outcomes are evaluated on an 'all or none' approach. We believe practice, and indeed research, may be better informed from a perspective that includes attention to progress toward treatment goals, mirroring a current trend toward the use of effect size estimates for evaluating treatment outcomes. With that goal in mind, we contend that, except for trend analysis, each of the methods presented provides a form of effect size estimate that allows for the determination of the *magnitude* of treatment outcomes, thereby providing data to indicate whether progress is being made toward a treatment goal. In practice settings, the magnitude of effect may lead to an evaluation of whether a given treatment should be discontinued, modified, or maintained.

R. T. Busse (✉) · K. S. Kennedy
Chapman University, Orange, CA, USA
e-mail: busse@chapman.edu

R. J. McGill
Texas Woman's University, Denton, TX, USA

Goal Attainment Scaling

Goal Attainment Scaling (GAS) (see Kiresuk et al. 1994) is a criterion-referenced rating scale approach that can be readily

applied in school and clinical settings. Kiresuk and Sherman (1968) developed the GAS method to evaluate the effectiveness of mental health services, using the technique to measure clients' progress towards treatment goals and to compare outcomes among different therapists and treatment programs. When applied in educational settings, multiple sources (i.e., teachers, aides, parents, consultants, or students) can complete GAS ratings of academic or social behavior intervention outcomes at individual or group levels. GAS has been found to be a reliable and valid method for evaluating intervention outcomes in school settings (Coffee and Ray-Subramanian 2009; Roach and Elliott 2005).

The basic GAS method involves selecting a target behavior, operationalizing behavioral or academic outcomes in objective terms, and ranking treatment effects from positive to negative outcomes (Elliott and Busse 2004; Roach and Elliott 2005). The outcomes are operationally defined on a five- or six-point rating scale, e.g., +2 to −2, wherein +2 indicates an intervention goal is attained, 0 indicates baseline, and −2 indicates that a problem is much worse. (Ratings from zero to 4 or 5 may be used for behaviors that have a basal rating of zero, such as a beginning reader learning the ABCs). By using ratings for each outcome level, a rater can provide hourly, daily, or weekly reports of student progress, depending on the target behavior. These ratings can be derived from direct indicators of progress (e.g., direct observations or permanent products) and/or from the raters' perceptions of a student's progress.

GAS goals may best be determined in a team format, wherein a teacher is involved in the process and, when appropriate, parents and students. The decision regarding goal definitions is idiosyncratic, in that the initial rating of zero is based on baseline data, and subsequent goals are based on individual student needs. Starting from baseline, the team uses professional judgment to set goals that are appropriate for a given student. For many behaviors, goals are defined as ranges. For example, a particular student may have a baseline of 45–55 % work completion in math. A GAS rating of 1 may be set at 56–89 % to indicate progress and a rating of 90 % or greater would indicate goal attainment, with mirrored negative ratings.

The strengths of the GAS method are that it (a) is time efficient; (b) can be used individually or for groups; (c) can be used as a self-monitoring tool; (d) can be used to repeatedly monitor progress; (e) is easy to graph and interpret; (f) can be used across settings and sources; (g) provides a method for evaluating the magnitude of progress toward outcome goals; and (h) is likely more readily understood by consumers than other quantified outcome methods, thereby lending social validity to the method. The limitations of GAS include (a) ratings for certain behaviors necessarily are based on summary observations and perceptions, such as goals for behaviors that may not lend themselves to accurate appraisal (e.g., teacher ratings of class participation), and (b) subjectivity is involved in deciding the level of goal attainment.

Visual Analysis

In single-case time-series outcome applications, the traditional method of analysis is visual inspection of graphed data. For changes in behavior from baseline to treatment to be indicative of improvement, the changes in level and variability between phases must be in the desired direction, generally immediate, readily discernible between phases, have separate trends, and be maintained across time.

There are several strengths of the visual analysis method (see Parsonson and Baer 1986). For example, visual analysis is quick to yield conclusions, graphing of data is relatively simple, data undergo minimal transformations, and the theoretical premises of graphs are minimal and easily understood. The limitations of visual analysis are that the method is potentially insensitive to subtle changes across baseline and treatment phases, difficulty in interpretation arises when there is variability and overlap of data or if the level or trend is unclear, and the graphing method may skew interpretation. Researchers have found that phase changes that are not readily discernible have resulted in inconsistent interpretation of graphed, time-series data (DeProspero and Cohen 1979; Knapp 1983).

Visual Analysis Rating

Visual Analysis Rating (VAR) has been proposed as a method to provide an augmentative quantitative index for visual analysis (Busse 2005). Drawing on the logic of goal attainment scaling, ratings are based on the degree of treatment response. The criteria upon which VAR is based are whether the change between baseline and treatment is in the desired direction with distinct trends, generally immediate, discernible, maintained, and whether there is a relatively large and stable average level of change. Ratings are 2=strong unequivocal evidence of the criteria; 1=data are in the desired direction, immediate, and discernible, but overlap between phases, and the average level change is moderate; 0=lack of evidence of any criteria; −1=same as 1, but in undesired direction; −2=same as 2, but in undesired direction.

McGill and Busse (2014) applied the VAR method in an assessment of outcomes for five elementary school children who received a remedial reading intervention. Each case was evaluated utilizing a mean difference effect size, VAR, and GAS ratings. VAR ratings tended to be more stable than the mean difference effect sizes and corresponded to the aggregate when the multiple evaluation methods were synthesized.

The strengths of VAR are that it is a relatively simple method, and it may be useful with intervention designs in which the ability to progress monitor is limited to a few data points. The limitations are that it is a relatively weak method, and it is in need of research on its reliability and validity.

Trend Analysis

The purpose of trend analysis is to examine the degree to which data are moving in a desired direction and toward a desired goal. The basic method for trend analysis is to create a regression line (a straight line of best fit for the data) for baseline and intervention data, and to analyze the slope of the line. When the trend line for the baseline phase is extended throughout the intervention period, it becomes possible to compare a student's progress to their trend in the absence of the intervention (Riley-Tillman and Burns 2009). In comparing trends, the same principles apply as with visual analysis; effective interventions will be represented with treatment phase data that trend in the desired direction and are distinct from baseline.

There are no available guidelines for determining when an observed change in slope from one phase to the other is significant. As our efforts focus on behaviors that may fluctuate without intervention (e.g., students who are behind in reading fluency may make weekly gains without additional services; behaviors may increase or decrease naturally based on changes in the environment), this comparison between a student's trajectories with and without the intervention allows us to be more conclusive that it is our intervention that influenced any gains.

Additionally, the slopes of trend lines can be used to inform us about the rates of growth that a student is making or is expected to make during an intervention. Weekly or daily growth rates can be used to evaluate progress and to assist in evaluating the feasibility of selected goals (e.g., if the student would be required to increase homework rate in all classes from a baseline of 50 to 100 % to reach the selected goal within a short timeframe, the goal may be shown to be unrealistic as indicated by the slope and trend). Another approach is to utilize established growth rates, such as those that are available for curriculum-based reading probes (e.g., Hasbrouk and Tindal 2006), to develop and plot a goal line for a fixed evaluation period. After several intervention sessions, a trend line can be calculated to determine whether the slope or trend is consistent with the goal line expectations.

Two methods of trend line calculation are widely used: the split-middle method and the ordinary least squares (OLS) approach (Kennedy 2005). The split-middle method utilizes the median data points from baseline and intervention phases to estimate trend lines and is easy to complete by hand (see Riley-Tillman and Burns 2009). OLS is a statistical regression method that provides a line of best fit to provide an estimate of direction and rate of improvement as indicated by the slope. OLS trend lines are easily created with basic graphic software programs.

Regression lines created via OLS are sensitive to outliers (Shinn et al. 1989) and may be inappropriate for cases in which data do not follow linear trends (e.g., a student's progress may follow a curved pattern as skills are acquired). OLS trend lines may be more accurate than those generated via the split-middle method (Shinn et al. 1989), and OLS is more widely used. In comparison to other evaluation methods, OLS requires a relatively large number of data points. Christ et al. (2012) recommended that a minimum of 14–15 data points may be needed before a reliable slope can be forecast with outcome measures such as CBM reading probes, although Parker et al. (2011) found that OLS predictions were valid with smaller data sets. Nevertheless, OLS continues to be recommended as a method to evaluate high-stakes single-case response to intervention decisions (Hixson et al. 2008).

It is important to note that OLS differs from the other methods discussed in this article, in that it cannot be interpreted as an effect size because it does not provide information relevant for estimating magnitude of change. Although the *R*-squared (R^2) statistic (which can be interpreted as an effect size) can be calculated from an OLS trend line with computer programs, most single-case research data do not meet the statistical assumptions for such analysis (Brossart et al. 2006).

The strengths of trend analysis are that it (a) is easy to complete with computer graphing programs; (b) allows us to account for trends that were in place before the intervention was implemented; (c) can be implemented with single-phase or multiple-phase intervention models; and (d) it is a powerful statistical method. The limitations of this approach are that (a) regression lines are less robust with highly variable data (Brossart et al. 2011); (b) OLS may require a relatively large number of data points for a robust outcome; and (c) there are no guidelines for quantifying the magnitude of differences between phase trends.

Percentage of Non-overlapping Data

Percentage of non-overlapping data (PND) provides an indicator of change in time-series data from baseline to treatment phases (Scruggs et al. 1987). The number of data points in the treatment phase that exceed the highest (or lowest if decreasing behavior) baseline data points are divided by the total number of data points in the treatment phase. A PND greater than or equal to 80 is indicative of a strong effect, 60–79 is moderate, and below 60 indicates no effect.

More sophisticated methods of the PND approach have been proposed, including the percentage of all non-overlapping data (Parker et al. 2007) and non-overlap of all pairs (NAP) (Parker and Vannest 2008). Whereas these methods, and in particular the PAND approach, have been cited as being more statistically robust (Riley-Tillman and Burns 2009), we believe the original method may be the most useful in practice due to the simplicity of calculation, and because it may be a more conservative effect size estimate.

The strengths of PND are that it is simple to compute, and it is relatively unaffected by nonlinearity and heterogeneity of the data. The limitations are that it is potentially oversensitive to a atypical baseline data (e.g., when baseline data are at extremes, such as zero for decreasing behaviors or 100 % for increasing behaviors for which there can be no overlap), it is adversely affected by trends, and it may not discriminate important treatment changes (White 1987). For example, an improvement from a high of 50 % work completion at baseline to a high of 53 % work completion after intervention is a strong effect with a PND of 100 % if the treatment phase data all surpass baseline, even though there was no discernible treatment effect.

Single-Case Mean Difference Effect Size Methods

There are several variations of mean difference effect size (MDES) methods. In this review, we focus on the method as an indicator of change from baseline to treatment in time-series data and on a variation that uses pre-post individual outcome data. Perhaps, the most popular and easy to compute method is the standardized mean difference (SMD) method for time-series data (Gingerich 1984). The SMD method is an extension of the logic of Smith and Glass (1977) and, later, Cohen's *d* (Cohen 1988) effect sizes for group designs to single-case research and is calculated by subtracting the mean of the treatment phase from the mean of the baseline phase and dividing the remainder by the standard deviation of either the baseline or the pooled phases. A minimum of 3 data points are necessary to compute the standard deviation (5 data points probably should be a minimum goal for each phase, although in practice this may not be feasible). In the case of no deviation (a completely stable baseline or treatment phase), one can approximate the effect size by using the phase standard deviation that demonstrates variability. Although variations of the SMD method exist (e.g., Gorman-Smith and Matson 1985), the SMD is most often calculated using all of the data points in each phase. The primary differences between methods regard assumptions about the distribution of variance across phases.

The no-assumptions approach (Busk and Serlin 1992) assumes the heterogeneity of variance between the treatment and baseline phases, and utilizes the standard deviation from the baseline phase. Another approach involves an assumption of homogeneity of variance across baseline and treatment phases, and uses the pooled within-phase variances (Busk and Serlin 1992). Some authors have stated that both baseline and intervention data must be normally distributed for the effect size to be interpretable (e.g., Riley-Tillman and Burns 2009). However, Busk and Marascuilo (1992) asserted that such assumptions are not possible due to design limitations unique to single-case research.

The SMD approach can be extended to a single-case score difference approach wherein an individual effect size is derived from pre-post data on some outcome measure for a group. In this approach, the pooled standard deviations of the group serve as the denominator and a Cohen's *d* is calculated for each individual. In this application, rather than analyzing the difference of the group mean, individual scores are computed.

The interpretation of SMD approaches typically is based on suggested guidelines for determining the magnitude of group design effect sizes (e.g., Cohen 1988), wherein 0.2 is considered a small effect, 0.5 a moderate effect, and 0.8 or larger a strong effect, and the effect is interpreted much like a *z* score. For example, an effect of 0.5 is interpreted as a half standard deviation change from baseline. Magnitude criteria for single-case outcomes have been infrequently applied in the literature and validated guidelines do not exist, thus these guidelines are in need of empirical validation, especially with the use of SMD effect size estimates. We suggest that the criteria for single-case data should be more conservative, with effect sizes of +1.0 (i.e., one standard deviation change) or greater indicating strong outcomes, effect sizes between +0.4 and +0.9 considered moderate outcomes, and effect sizes less than 0.4 indicating no appreciable treatment effect or a mirrored negative effect.

The time-series and individual score differences effect size approaches are strong inclusions in the assessment of single-case outcomes for several reasons. First, the SMD method provides data that can be used to quantify the effectiveness of a particular intervention, which is congruent with the growing evidence-based practice movement in school psychology (e.g., Kratochwill and Stoiber 2002). Second, the method is easy to compute with available Internet-based programs, and it represents one of the more robust single-case effect size estimates that are accessible to practicing school psychologists.

The strengths of the method are that it provides a means for quantifying time-series single-case outcomes, the effect size can be interpreted much like a *z* score, and it is a relatively robust effect size method. The limitations are that the effect size estimates do not account for trend or autocorrelation, and phase means may be affected by outliers.

Reliable Change Index

Reliable Change Index (RCI) is a method for measuring the difference between pretest and posttest scores on a dependent measure, usually a standardized test (Jacobson et al. 1984; Jacobson and Truax 1991). The simplest method is to subtract the two scores and divide by the standard error of measurement (SEM). More sophisticated methods are to use the standard error of the estimate or the standard error of difference (these methods are more conservative in that

they increase the denominator in the equation and render a smaller effect).

Although several RCI equations have been proposed (e.g., Hageman and Arrindall 1999; Hsu 1989), the most popular and often utilized is the Jacobson and Truax (1991) formula. According to this method, a pretest baseline score from an outcome measure is subtracted from the posttest score, and the result is divided by the standard error of difference of the outcome measure.

This RCI equation is inextricably linked to the reliability of the outcome measure that is utilized. Less reliable instruments create a larger error term, thus larger differences between pretest and posttest scores are needed for the difference to be considered significant. Conversely, smaller differences are needed for significance when more reliable instruments are used. RCI values that exceed 1.96 are considered to be statistically significant. This critical value is not arbitrary as it corresponds to a two standard deviation improvement from the baseline score.

Busse and colleagues (Busse 2005; Busse et al. 2010; Busse and Yi 2013; Elliott and Busse 2004) suggested that RCI can be interpreted as an effect size with regard to magnitude and proposed the following guidelines: RCIs$\geq$1.8 are indicative of a strong, positive change, RCIs from 0.7 to 1.7 are indicative of moderate change, −0.6 to 0.6 are indicative of no behavioral change, −0.7 to −1.7 are indicative of a moderate negative effect, and RCIs$\leq$−1.8 indicate that a behavior problem has significantly worsened. These are only guidelines and are in need of validation.

The RCI is a potentially useful measure of clinical change that is readily available to practitioners. It has been demonstrated to be more reliable than clinical judgment and client self-reports in accounting for meaningful change in counseling settings (Lunnen and Ogles 1998). The RCI is fairly easy to calculate (not taking into account the more complex regression methods); the needed values of reliability and SEM of the outcome measure are available in the manuals of most commercial instruments or can be calculated from reliability estimates provided for research-based measures. Whereas some calculations are needed, the technical sophistication required is on par with other single-case effect sizes that are often utilized in behavioral research, and there are available Internet-based calculation sites. RCI may be particularly well suited for behavioral response to intervention methods that utilize rating scale technologies for screening and progress monitoring.

Whereas the RCI has many positive attributes, it is not without its shortcomings. Although the RCI can be interpreted with regard to magnitude of effect, this use may eliminate properties that may be useful for clinical validation (Tingey et al. 1996). For example, a difference magnitude may be indicative of a significant change; however, a student may still be functioning within an impaired range. The statistic also is dependent on the utilization of outcome measures that demonstrate adequate reliability for educational and psychological decision making. Many outcome measures may lack treatment sensitivity (e.g., standardized achievement tests) and other basic psychometric properties that are needed to validly utilize the RCI.

The strengths of RCI are that it can be interpreted much like a *z* score, reliable changes are indexed by a standard error term, confidence intervals can be constructed around it, and it can be used to determine the magnitude of an effect. The limitations are that it is sensitive to the reliability of the instruments utilized, and it is limited to pre-post designs.

Convergent Evidence Scaling

Convergent Evidence Scaling (CES) is an emergent method designed to aggregate assessment data from multiple indexes for the evaluation of intervention outcomes (Busse et al. 2010). Multiple assessment data traditionally are "converged" through visual inspection and decisions about the magnitude of the effects or findings, without a systematic process for reaching conclusions about treatment outcomes. CES systematizes outcome analysis through the logic of GAS, such that ratings for each outcome index are used to provide a common metric for quantifying the data. Each outcome index is operationalized and converted into a common scale, which allows for combining data from multiple sources and methods into a single quantified index of the magnitude of change toward intervention goals.

The basic elements of the CES method are a five-point scale and definitions of outcome magnitudes that correspond to the following outcomes: Treatment goal fully met: strong positive effect (+2); treatment goal partially met: moderate positive effect (+1); no progress toward goal: no effect (0); treatment goal unmet, behavior somewhat worse: moderate negative effect (−1); and treatment goal unmet, behavior significantly worse: strong negative effect (−2).

For example, consider an adapted composite case example from Busse et al. (2010) based on teacher ratings on a narrowband aggression scale and GAS, and a school psychologist's observation of aggressive behavior at baseline and during 3 weeks of treatment. The data from the teacher rating scale (M=50; SD=10) resulted in an elevated standard score (72) at pretest and an average range score (60) at posttest. This change yielded a RCI of +3.0, which is then converted into a CES rating of +2. The teacher's mean GAS rating is +0.44, which is converted into a CES rating of 0. The school psychologist's observations of aggressive behavior yielded a single-case mean difference effect size of 1.8, which is converted into a CES rating of +2. The mean of the individual CES ratings is 1.33, which is converted into an overall CES outcome rating of +1 (moderate positive effect).

Busse et al. (1995) used the CES method to evaluate behavioral consultation outcomes in a meta-analytic framework, and later as a dependent variable in a related multiple regression analysis (Busse et al. 1999). McGill and Busse (2014) used the CES method to evaluate the outcomes from a small group reading fluency intervention and found that it was more consistent than other evaluation methods at discriminating between adequate and inadequate response to treatment.

There are several potential advantages of the CES method. First, it allows for converging data into an overall quantity from indexes that have different methods of interpretation, somewhat akin to converting raw score data from different scales into standard scores. The CES method transforms all outcome data into an ordinal scale, which allows for converging data from different types of outcome measures, including those utilizing interval level data. In research, the CES method can be used within a meta-analytic study to compare treatment outcomes from studies that use different methods of data collection (e.g., rating scales and time-series data), and the method can be used within a given study to combine treatment outcomes. Other strengths of the method are that CES relies on a criterion-referenced approach for assessing intervention outcomes, which has the potential for being a better method of measuring clinical or educational significance and may be most useful with single-case methods, and CES allows for the assessment of intermediate progress toward intervention goals.

A limitation of CES arises when outcomes are mixed (e.g., GAS indicates a strong effect and RCI indicates no effect), which are averaged in the overall CES rating and may obscure useful data, such as in situations where higher, multiple CES ratings attenuate a lower rating. A related limitation is the number of assessment indexes used; one may gather GAS ratings, observations, and rating scale data from multiple sources, all of which would be included in the CES ratings. If the majority of CES ratings are high or low, an outlier that potentially provides useful data will be obscured, or using fewer indexes may skew the relative weight of a single high or low treatment outcome. Other limitations include the potential for increased Type I error (accepting a finding as 'true' when it is not) due to multiple testing, and the method is in need of further validation.

Data Synthesis Display for Single-Case Outcomes

In this section, we provide a multiple index outcome data display that may facilitate decision-making processes (see Fig. 1). The display is an extension of a method proposed by Busse et al. (1995). To demonstrate the evolution of single-case outcomes, we use the data from Busse et al. but changed one case to work completion and augmented both cases with fictional data to illustrate the use of the outcome assessment methods presented in this article for behavioral and academic problems.

As shown in Fig. 1, the data display includes brief contextual information on the target behavior and intervention, along with various outcomes. As shown in Case 1, visual analysis indicates a change from baseline to treatment with a relatively large level change. The trend analysis indicates that the baseline data were trending in the same direction as the treatment phase; therefore, one cannot conclude that the observed increase in work completion was due to the treatment. The VAR of +1 reflected this conclusion. The mean GAS rating of +1 for the treatment phase indicates a moderate positive effect. The PND is 75 % which indicates a moderate positive effect, and the MDES is +1.97, indicating a strong positive effect. The CES individual outcome conversions are +1 for all but the MDES, which is a CES of +2. The mean individual CES is +1.2, which is converted to an overall CES of +1, indicating an overall moderate, positive effect.

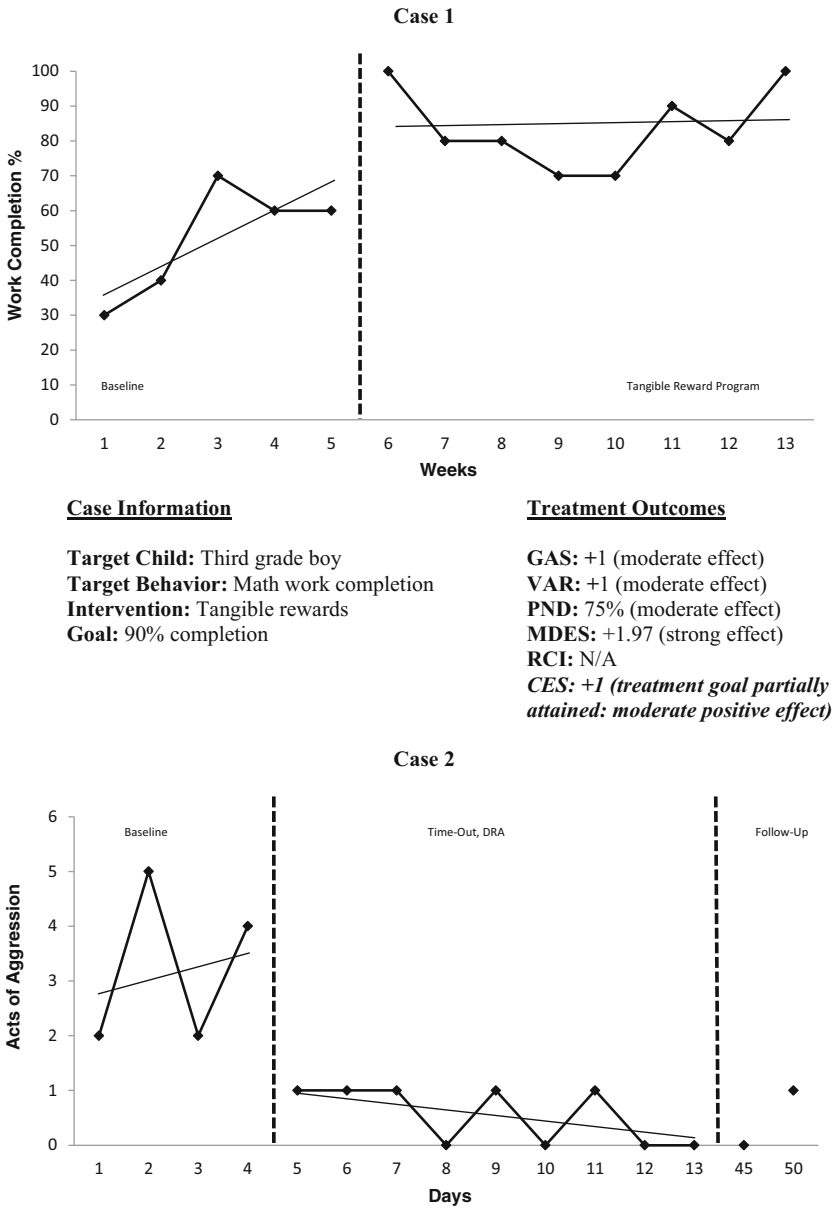
For Case 2, visual analysis indicates a relatively large level change that was maintained at a 1-month follow-up, and the trend analysis indicates the phases are discernible, which resulted in a VAR of +2, or a strong positive effect. The PND is 100 %, the MDES is +1.81, and the mean GAS rating for the treatment phase is +2, all of which indicate a strong positive effect. The CES individual outcome conversions were all +2, with a mean individual CES of +2, which is converted to an overall CES of +2, indicating an overall strong positive effect.

Implications for Practice

The methods we describe and the data display have several potential practical applications. Practitioners can use the methods for data-based decision making when evaluating behavioral and academic intervention outcomes. The methods can be used to demonstrate to administration and other stakeholders whether the interventions we use in school are effective and to provide a quantitative synthesis of data for record keeping purposes.

We chose several effect size methods in lieu of others. Our choices were purposeful: we favored methods that may be more readily understood and utilized in school-based practice. Although researchers may use more statistically sophisticated methods, the typical practitioner likely will not have the time, training, or resources to engage in their use (and all the methods have strengths and limitations). The methods we presented are easily calculated either by hand or with available online resources. A very useful online resource of which many practitioners may be aware is www.interventioncentral.org. The website contains a program called "ChartDog" that provides a time-series graph and displays an OLS trend line with slope data, PND, and a single-case means difference

Fig. 1 Data synthesis display for single-case outcomes



effect size. These data can be easily incorporated into a computer program to produce the graph and data display such as the one we present that was created on EXCEL. The choice of which method(s) to use/include in the display to evaluate a treatment outcome(s) depends of course on the type of data gathered.

Discussion and Conclusion

The purpose of this article is to present the use of multiple single-case outcome methods for evaluating treatment effectiveness. Several issues arise from this approach. One issue is encompassed in the concept of *evaluation integrity* (or

fidelity). It has become a basic and necessary understanding in research and practice that treatment integrity is integral to evaluating whether our interventions evidence internal validity (Gresham 1989). The concept of evaluation integrity, however, is infrequently broached although it is perhaps no less important in our appraisal of research and practice based outcomes. It is of paramount importance that we monitor treatment outcomes in a systematic format. Otherwise, how do we know that even well-designed treatments implemented with integrity have been appropriately evaluated? Another issue that is intertwined with this discussion is the reliability and validity of the measures selected to monitor progress and makes decisions about the effectiveness of interventions. Further research to establish or improve the psychometric properties of academic and behavioral progress monitoring tools should be a continued priority in our profession.

Other issues relate to the use of quantified effect sizes for assessing single-case outcome data. Several authors have eschewed the quantification of single-case outcomes (e.g., Baer 1977; Michael 1974; Parsonson 2003), or have stated that single-case effect sizes may be best used as an accompaniment for traditional visual analysis (e.g., Riley-Tillman and Burns 2009). Both of these stances have merit, in that there is a loss of qualitative information through the quantification of data, and each effect size method has significant limitations that must be considered in the decision making process. Furthermore, the methods and interpretations we described are in need of empirical validation.

The debate surrounding the use of effect sizes in single-case designs is healthy. As we advance toward further validation of these methods, we anticipate their use will evolve to meet technical and clinical advances in outcome analyses. At this point, in the evolution of single-case outcome methods, there is no consensus regarding the best approach to evaluating intervention outcomes. The inclusion of multiple outcome assessment methods may serve to offset the inherent limitations of each approach, much as best practices in problem assessment takes into consideration variance across multiple methods, settings, sources, and time. Our primary goal was to present methods that may be useful in practice applications and decision making processes. To that end, we invite further discussion and debate.

References

Baer, D. M. (1977). Perhaps it would be better not to know everything. *Journal of Applied Behavior Analysis, 10*, 167–172.

Brossart, D. F., Parker, R. I., Olson, E. A., & 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.

Brossart, D. F., Parker, R. I., & Castillo, L. G. (2011). Robust regression for single-case data analysis: How can it help? *Behavior Research Methods, 43*, 710–719.

Busk, P. L., & Marascuilo, L. A. (1992). Statistical analysis in single-case research: Issues, procedures, and recommendations, with applications to multiple behaviors. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 159–186). Hillsdale: Erlbaum.

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

Busse, R. T. (2005). *Methods for assessing response to intervention*. Atlanta: Workshop presented at the annual convention of the National Association of School Psychologists.

Busse, R. T., & Yi, M. (2013). Behavioral and academic rating scale applications within the problem-solving model. In R. Brown-Chidsey & A. Andren (Eds.), *Assessment for intervention: a problem-solving approach* (2nd ed., pp. 180–198). New York: Guilford.

Busse, R. T., Kratochwill, T. R., & Elliott, S. N. (1995). Meta-analysis for single-case consultation outcomes: Applications to research and practice. *Journal of School Psychology, 33*, 269–285.

Busse, R. T., Elliott, S. N., & Kratochwill, T. R. (1999). Influences of verbal interactions during behavioral consultations on treatment outcomes. *Journal of School Psychology, 37*, 117–143.

Busse, R. T., Elliott, S. N., & Kratochwill, T. K. (2010). Convergent evidence scaling for multiple assessment indicators: Conceptual issues, application, and technical challenges. *Journal of Applied School Psychology, 26*, 149–161.

Christ, T. J., Zopluoglu, C., Long, J. D., & Monaghen, B. D. (2012). Curriculum-based measurement of oral-reading: Quality of progress monitoring outcomes. *Exceptional Children, 78*, 356–373.

Coffee, G., & Ray-Subramanian, C. (2009). Goal attainment scaling: a progress-monitoring tool for behavioral interventions. *School Psychology Forum, 3*, 1–12.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Psychology Press.

DeProspero, A., & Cohen, S. (1979). Inconsistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis, 12*, 573–579.

Elliott, S. N., & Busse, R. T. (2004). Assessment and evaluation of students' behavior and intervention outcomes: the utility of rating scale methods. In R. B. Rutherford, M. M. Quinn, & S. R. Mathur (Eds.), *Handbook of research in emotional and behavioral disorders* (pp. 123–142). New York: Guilford.

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

Gorman-Smith, D., & Matson, J. L. (1985). A review of treatment research for self-injurious and stereotyped responding. *Journal of Mental Deficiency Research, 29*, 295–308.

Gresham, F. M. (1989). Assessment of intervention integrity in school consultation and prereferral interventions. *School Psychology Review, 18*, 37–50.

Hageman, W. J., & Arrindell, W. A. (1999). Establishing clinically significant change: Increment of precision and the distinction between individual and group level of analysis. *Behavior Research and Therapy, 37*, 1169–1193.

Hasbrouk, J. E., & Tindal, G. A. (2006). Oral reading fluency norms: a valuable assessment tool for reading teachers. *The Reading Teacher, 59*, 636–644.

Hixson, M., Christ, T. J., & Bradley-Johnson, S. (2008). Best practices in the analysis of progress monitoring data and decision making. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 2133–2146). Bethesda: National Association of School Psychologists.

Hsu, L. M. (1989). Reliable changes in psychotherapy: Taking into account regression toward the mean. *Behavioral Assessment, 11*, 459–467.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12–19.

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy, 15*, 336–352.

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

Kiresuk, T. J., & Sherman, R. E. (1968). Goal attainment scaling: a general method for evaluating community mental health programs. *Community Mental Health Journal, 4*, 443–453.

Kiresuk, T. J., Smith, A., & Cardillo, J. E. (Eds.). (1994). *Goal attainment scaling: Application, theory, and measurement*. Hillsdale: Erlbaum.

Knapp, T. J. (1983). Behavior analysts' visual appraisal of behavior change in graphic display. *Behavioral Assessment, 5*, 155–164.

Kratochwill, T. R., & Stoiber, K. C. (2002). Empirically supported interventions and school psychology: Conceptual and practice issues: Part II. *School Psychology Quarterly, 15*, 233–253.

Lunnen, K. M., & Ogles, B. M. (1998). A multiperspective, multivariable evaluation of reliable change. *Journal of Consulting and Clinical Psychology, 66*, 400–410.

McGill, R. J., & Busse, R. T. (2014). An evaluation of multiple single-case outcome indicators using convergent evidence scaling. *Contemporary School Psychology, 18*, 13–23.

Michael, J. (1974). Statistical inference for individual organism research: Mixed blessing or curse? *Journal of Applied Behavior Analysis, 7*, 647–653.

Parker, R. I., & Vannest, K. J. (2008). An improved effect size for single-case research: nonoverlap of all pairs. *Behavior Therapy, 40*, 95–105.

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

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy, 42*, 284–99.

Parsonson, B. S. (2003). Visual analysis of graphs: Seeing is believing. In K. S. Budd & T. Stokes (Eds.), *A small matter of proof: the legacy of Donald M. Baer* (pp. 35–52). Reno: Context Press.

Parsonson, B. S., & Baer, D.M. (1986). The graphic analysis of data. In A. Poling & R.W. Fuqua (Eds.), *Research methods in applied behavior analysis: Issues and advances* (pp. 157–186). New York: Plenum.

Riley-Tillman, T. C., & Burns, M. K. (2009). *Evaluating educational interventions: Single-case design for measuring response to intervention*. New York: Guilford.

Roach, A. T., & Elliott, S. N. (2005). Goal attainment scaling: an efficient and effective approach to monitoring student progress. *Teaching Exceptional Children, 37*(4), 8–17.

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

Shinn, M., Good, R., & Stein, S. (1989). Summarizing trend in student achievement: a comparison of methods. *School Psychology Review, 18*, 356–370.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32*, 752–760.

Tingey, R. C., Lambert, M. L., Burlingame, G. M., & Hansen, N. B. (1996). Assessing clinical significance: Proposed extensions to the method. *Psychotherapy Research, 6*, 109–103.

White, O. R. (1987). Some comments concerning "The quantitative synthesis of single-subject research.". *Remedial and Special Education, 11*, 203–214.

R.T. Busse, PhD is an Associate Professor in the Counseling and School Psychology Program at Chapman University. Dr. Busse's interests include direct assessment of academic skills, single-case methods for assessing intervention outcomes, and selective mutism.

Ryan J. McGill, PhD is an Assistant Professor in the Department of Psychology and Philosophy at Texas Woman's University. His research interests include data-based decision making in school psychology, behavior analysis and intervention, and single-case methods.

Kelly S. Kennedy, PhD is an Associate Professor in the Counseling and School Psychology program at Chapman University. Her research focuses on improving school-based practices in many areas, including multicultural competence, counseling, and data-based decision making. Dr. Kennedy is a member of the Consortium for the Advancement of School Psychology in Vietnam (CASP-V), is the editor for *Trainers' Forum*, and an associate editor of *Contemporary School Psychology*.