

Effect Sizes in Single Case Research: How Large is Large?

Richard I. Parker, Daniel F. Brossart, and Kimberly J. Vannest
Texas A&M University

James R. Long
VA Palo Alto Health Care System

Roman Garcia De-Alba
Cypress-Fairbanks ISD

Frank G. Baugh
National Center for Organization Development

Jeremy R. Sullivan
Cypress-Fairbanks ISD

Abstract. This study examined the problem of interpreting effect sizes in single case research. Nine single case analytic techniques were applied to a convenience sample of 77 published interrupted time series (AB) datasets, and the results were compared by technique across the datasets. Reanalysis of the published data helped answer questions about the nine analytic techniques: their effect sizes, autocorrelation, statistical power, and intercorrelations. The study's findings were that few effect sizes matched Cohen's (1988) guidelines, and that effect sizes varied greatly by analytic technique. Four techniques showed adequate power for typical published data series. Autocorrelation was a sizeable problem in most analyses. In general, individual techniques performed so differently that users need technique-specific information to guide both selection of an analytic technique and interpretation of its results.

The debate on the usefulness of statistical analysis with single case research data has largely been resolved over the past decade. Though it is acknowledged that no present statistical technique can adequately reflect the range of criteria available to visual analysis (Baer, 1977; Michael, 1974; Parsonson & Baer, 1992), statistical analysis is now regarded by most experts as a useful supplementary technique in many circumstances. Even strong proponents of visual analysis (Huitema, 1986;

Kazdin, 1982) acknowledge that statistical results can be valuable or even essential when there is no stable baseline, when unambiguous results must be shared with other professionals, and when effects of a new treatment cannot be predicted. Interestingly, the first two of these three conditions are common. Although stable, flat baselines are desirable, they are often not found—even in published data. Of 77 published graphs composing the convenience sample for this study, nearly 66% had

Correspondence concerning this article should be emailed to Richard I. Parker, at rparker@tamu.edu

Copyright 2005 by the National Association of School Psychologists, ISSN 0279-6015

noticeable positive or negative baseline trend, and over 50% of the baselines possessed high variability.

The second condition arguing for statistical analyses, the need to document and share data unambiguously, is commonplace with shrinking external funding, and increased accountability for use of those funds. Funding agents increasingly require evaluation of client interventions to establish treatment efficacy through objective, quantifiable data. Although visual analysis conclusions are convincing to individual clinicians, evidence converging from multiple studies informs one that visual judgments of graphed data are notoriously unreliable. Finally, the application of meta-analysis to single case research has brought to focus the need for valid, objective measures of treatment effects that can be communicated beyond the walls of a particular clinical context, and compared with results from other environments.

The acknowledged benefit of statistical analysis as a supplementary technique in many or most circumstances has not, however, translated to its broad use in published data. In this study's sample of 124 articles from counseling, clinical, and school psychology journals over the past 15 years, over 65% used only visual analysis. Nonstatistical comparisons of means, medians or proportions comprised most of the remaining 35%. Effect sizes, confidence intervals or tests of statistical significance were found in only 11% of the 124 articles. This prevalence is comparable to the 10% prevalence rate for statistical analyses found in earlier, larger surveys 10 and 25 years ago (Busk & Marascuilo, 1992; Kratochwill & Brody, 1978).

The underuse of statistical analysis of single case research data is the context for the present study. This underuse is understandable, as researchers or clinicians wishing to supplement visual with statistical analyses have available a number of techniques, but little information on how any of them performs. The number of analytic techniques available for short data series has easily tripled since the early 1980s (Barlow & Hersen, 1984; Kazdin, 1982), yet promising techniques such as the

regression models of Center et al. (Center, Skiba, & Casey, 1985-1986) and of Allison and colleagues (Allison & Gorman, 1993; Faith, Allison, & Gorman, 1996) have to date been applied and examined in only a small handful of studies. They are known mainly through summaries in texts by Franklin, Allison, and Gorman (1996) and Kratochwill and Levin (1992).

The underuse of statistical analyses by single case researchers is also certainly due to concerns about autocorrelation or serial dependence among the time series data. Autocorrelation violates assumptions of most statistical techniques (Fox, 1991), and exists at problematic levels in from 83% (Jones, Vaught, & Weinrott, 1977; Suen & Ary, 1987) to less than 30% (Hartmann et al., 1980; Huitema, 1985) of published single case data. Even small autocorrelation levels of $r = .20-.30$ are said to increase Type I error rate by a factor of 2 to 3 (Scheffé, 1959). Because autocorrelation is calculated on residual scores, and because residuals are the product of the interaction between the predictors and criterion scores in a regression model, it is logical that autocorrelation varies by analytic model (Gorman & Allison, 1996). The limited evidence available indicates that this holds in single case research (Busk & Marascuilo, 1988; Center et al. 1985-1986; Holtzman, 1963; Kazdin, 1984). Center et al. (1985-1986) found that autocorrelation in their analytic models was greatly reduced (by over .40) from using a standard ANOVA analysis. These limited findings underscore the need to examine and compare single case analytic techniques for the autocorrelation levels they tend to produce. Clinicians and researchers should be forewarned about such differences so they can make informed decisions in selecting an analytic technique.

The comparative performance of single case statistical techniques has been little explored. The few comparative studies have concluded that different techniques produce quite different results. Nourbakhsh and Ottenbacher (1994) found that three supposedly similar statistical indices (Tryon's *C* statistic, two-standard deviation band method, Owen White's

split-middle method) performed very differently on the same data series.

Consequently, researchers lack normative bases for interpreting effect sizes from any given analytic technique. Do the various available techniques yield comparable or quite different effect sizes? To what extent do the various techniques tend to agree with one another (covary) in analysis of the same data? What is the statistical power of these techniques to detect noteworthy effects in the relatively short data sets available to most single case clinicians and researchers? To what extent is autocorrelation (serial dependency) an issue with each technique? Answers to such questions are needed for scientist-practitioners to use the statistical techniques comfortably.

The effect size offers several advantages as a summary of behavior change across phases. First, its focus is strength of association—the extent to which a response variable can be explained, predicted, and controlled by the intervention (Carver, 1978; Mitchell & Hartmann 1981; Rosnow & Rosenthal, 1989). Given a design with sufficient internal validity, an effect size index also indicates degree of treatment success. Because the effect size is continuously scaled, it can support incremental treatment decisions. Finally, effect sizes are only indirectly affected (through reduced distribution variability) by the small sample sizes common in single case research. Busk and Serlin (1992) conclude that an effect-size measure is the “obvious choice” for quantifying single case research results (p. 192).

Of the numerous effect size indices available, only two— R^2 and Eta-squared—are common in APA journals (Kirk, 1996). Any effect size can be transformed to any other through simple mathematical formulae (Cohen, 1988; Rosenthal, 1991). In single case research, a second common effect size family is the “standardized mean difference” (Cohen’s d , Glass’s g , Hedges g) (Hedges & Olkin, 1985). Cohen’s d is well suited to tests of simple mean difference between phases. However, analysts using complex models that include trend have favored R^2 for its greater flexibility in interpreting single case data. After reviewing most available analytic techniques for single case

data, Franklin et al. (1996) concluded that regression approaches are imperfect, but the best available.

Cohen (1988) has provided the only widely accepted guidelines for interpreting effect sizes, with anchors for “large” ($R^2 = .25$), “medium” ($R^2 = .09$) and “small” ($R^2 = .01$) effects, based on continuous predictors, and for categorical predictors (with a point-biserial distribution): “large” ($R^2_{\text{diff}} = .137$), “medium” ($R^2_{\text{diff}} = .059$), and “small” ($R^2_{\text{diff}} = .01$) (p. 82). However, Cohen stressed that these guidelines were derived from large group social science research, and may not fit other types of research. Kirk (1996) similarly cautioned that given the contextual dependency of effect sizes, one should not overgeneralize any guidelines. Several other authors have warned about the influence of design, client, and intervention differences on effect size magnitudes (Maxwell, Camp, & Avery, 1981; Mitchell & Hartmann, 1981; Rosnow & Rosenthal, 1989). In addition, it can be expected that mean difference versus mean plus trend difference analytic models will influence the magnitude of effects, independently of intervention effectiveness. In addition, the magnitude of effects is likely influenced by whether and how trend is controlled (Cohen, 1988).

Statisticians (Fidler & Thompson, 2001) now recommend that effect size presentation include an index of reliability such as confidence intervals (CIs; Fowler, 1985). Little is presently known about the typical reliability of results from single case analytic techniques (Allison, Silverstein, & Gorman, 1996). Though the single case research literature is replete with warnings against insufficient data points, these cautions lack specificity.

For a large number of analyses, as conducted in the present study (77 datasets \times 9 techniques = 558), an alternative approach to assaying effect size reliability is to conduct power analyses, which inform the researcher what critical effect size levels must be reached to obtain statistical significance for a given sample size and type of regression analysis (Cohen, 1988). Power analyses for a range of critical effect sizes yields a power graph that helps estimate whether the number of obser-

variations is likely to be sufficient. In this study, power graphs were plotted, although with one or a few clients, confidence intervals are preferred. Confidence intervals are based on the asymmetrical F distribution, and can be calculated with the freeware R^2 (Steiger & Fouladi, 1992).

Method

Selection of Published Data

The single case datasets for this study were from 77 graphs within 26 published articles from 13 different journals. The datasets were found through ERIC and PsycINFO searches covering the past 20 years, using search terms “single case,” “single subject,” “time series,” “baseline,” and “AB,” “ABA,” “ABAB,” and “ABC.” Graphs had to be large and clear enough for digitizing, and had to permit an AB comparison. Only AB phase comparisons were analyzed in this study, although any other common contrast could have been selected. Selection criteria were: (a) presence of baseline and intervention phases, (b) a minimum of 6 data points per phase, and (c) at least 14 data points in phases A and B together, and (d) graphs large and clear enough for scanning. These requirements are more stringent than in most previous studies (Jones et al., 1977; Matyas & Greenwood, 1996), though similar to the study by Huitema (1985). The more stringent selection criteria for this study was because of the present focus on statistical, rather than visual, analysis.

This study used a convenience sample. From the ERIC and PsycINFO searches, 124 promising articles were obtained. By chance alone, *the Journal of Applied Behavior Analysis* (JABA) was poorly represented (only one article), so five additional JABA articles were added, for a total of 129 articles, containing 362 graphs. The four selection criteria identified 77 useable graphs within 26 articles, located in 13 different journals. Most unusable graphs had too few data points.

Most (18) of the chosen articles contributed 1 to 3 graphs each, but 6 articles contributed 4 graphs each, and 2 articles contributed 5 graphs each. Of the 26 articles, 8 offered

multiple-baseline designs, 9 ABC designs, 8 AB designs, and 1 ABAB design. The AB designs were mainly from counseling or clinical psychology, and the multiple-baseline designs were mainly from school psychology or special education. The 26 articles sampled are included in the References section, indicated with an asterisk. The median number of data points per graph (counting only A and B phases) was 23, with an interquartile range (IQR) of 18 to 30. For phase A, the median length was 10, and the IQR was 7 to 14. For phase B, the median was 11, with the IQR 9 to 16.

Published graphs were digitized using *i-extractor* software (Linden Software Ltd., 1998), following four steps. First, graphs were scanned at 300 dpi into a computer, and the jpg files opened with *i-extractor*. Graph axes were set to provide actual data values on a digital Cartesian coordinate spreadsheet. Clicking on each data point then read its value into an Excel spreadsheet. Data values were finally regraphed, and these graphs compared with the originals from the articles. Original and recreated graphs were compared by sizing the new graphs to the same physical dimensions of the original graphs. Then the two were stacked and held against a bright window. Exact overlap of data points was required before proceeding with analyses. A few erroneous graph points were identified through this method, and corrected.

Nine Analytic Techniques

The nine analytic techniques investigated span over 25 years of single case research. The popular texts by Kazdin (1982) and Barlow and Hersen (1984) were relied on for earlier techniques. For more recent techniques, texts by Kratochwill and Levin (1992) and by Franklin, Allison, and Gorman (1996) were used.¹ Mean-only and Mean plus Trend models were included, but not Trend-only models, as improvement in trend but not in mean level is generally not an accepted indicator of an effective intervention.

Effect sizes from these measures were to reflect client improvement, not merely client change. Effect sizes are blind to improvement versus change, so visual analysis was used

to detect any cases of deterioration from phase A to B. The R^2 for deteriorated trend in a mean+trend difference model reverted to a simple mean difference calculation. The R^2 for deteriorated mean differences reverted to zero, because negative effect sizes are not possible. Table 1 describes each technique and the number of cases needing adjustments for phase B deterioration.

Graphs were scanned, their data were digitized and saved into an Excel spreadsheet, and then transferred to Number Cruncher Statistical Software (NCSS; Hintze, 2002), which includes both time series and power analysis modules. All nine statistical analyses were then conducted on each of the 77 datasets, and results were input to a summary data spreadsheet of effect sizes and autocorrelation coefficients.

Table 1
Nine Statistical Analysis Techniques for Single-Case Data

Analytic Technique	Description
SIMP-M: Simple mean shift test (Cohen & Cohen, 1983).	Client scores are regressed on a dummy-coded (0/1) Phase variable. The effect size is the resulting R^2 . No SIMP-M results required adjustment for deteriorating Phase B performance.
FULL-MT: Full Model, or Mean Plus Trend Interaction Model (Cohen, 1988).	FULL-MT includes main effects for both mean level and trend, as well as their interaction. Client scores are regressed on Trend (the time variable) and Phase (dummy-coded 0/1) predictors in a multiple regression, with an interaction term. Eight of the FULL-MT results showed negative Phase B trend so that negative trend was eliminated from the analysis.
BINOM: binomial test on extended Phase A baseline (White & Haring, 1980).	In BINOM, the Phase A median slope is hand-fit to evenly split the Phase A data (50% above and below the line). That Phase A line is then extended through Phase B, and a binomial test performed on the splitting of the Phase B data (Darlington & Carlson, 1987) by the extended Phase A trend line. The resulting Z score with continuity correction was converted to an r^2 effect size: $R^2 = Z^2/N$ (Rosenthal, 1991). Nine of the BINOM results showed deterioration in the intervention phase, so their effect sizes were adjusted to zero.
LTD: Last Treatment Day (White et al., 1989).	LTD compares performance levels predicted at the end of the treatment phase from two different regression lines—from an extended Phase A regression line, and from the Phase B regression line. These two predicted values are subtracted and the difference is divided by standard error of prediction error term (Nunnally, 1978) to obtain Cohen's d .
$d = \frac{LTD_B - LTD_A}{\sqrt{SD_{Pooled}(1 - r^2)}}$	
	The d is then converted to an R^2 effect size through: $R^2 = d^2/(d^2 + 4)$ (Rosenthal, 1991). Ten of the LTD results showed deterioration during treatment, so their R^2 was converted to zero.
GORS: Gorsuch's trend analysis effect size (Faith et al., 1996; Gorsuch, 1983).	GORS tests mean differences between phases, while controlling for overall data trend. First, the entire data series is detrended; i.e., trend is semipartialled from scores. Then the detrended scores are regressed on a dummy-coded (0,1) phase vector. No adjustments for deteriorated Phase B mean levels were required for GORS analyses.

(Table 1 continues)

(Table 1 continued)

Analytic Technique	Description
CENT-M: Center Mean-only Model (Center et al., 1985-1986; Berry & Lewis-Beck, 1986; Kromrey & Foster-Johnson, 1996).	<p>CENT-M tests for between-phase differences while controlling for overall data trend. Trend is fully partialled, not semipartialled:</p> $\frac{R^2_{Y \cdot TM} - R^2_{Y \cdot T}}{1 - R^2_{Y \cdot T}}$ <p>(<i>Y</i> = client response, <i>T</i> = time or linear trend, <i>M</i> = 0/1 phase vector). This study required a more elaborate calculation method (Cohen, 1983, Chapt. 3) to obtain autocorrelation output. No CENT-M results indicated deteriorating Phase B results, so no effect size adjustments were required.</p>
CENT-MT: Center Mean plus Trend Model (same citations as for CENT-M).	<p>CENT-MT tests for combined mean and trend differences between baseline and intervention phases while controlling for overall data trend (trend fully partialled out):</p> $\frac{R^2_{Y \cdot T, TM} - R^2_{Y \cdot T}}{1 - R^2_{Y \cdot T}}$ <p>(<i>Y</i> = client response, <i>T</i> = time or linear trend, <i>M</i> = 0/1 phase, <i>TM</i> = interaction term). This study used a more elaborate method (see Cohen, 1983, Chapt. 3) to obtain residual autocorrelation output. For CENT-MT, 24 results showed Phase B deterioration, so the negative Phase B trend was neutralized.</p>
ALLIS-M: Allison et al.'s mean-only model (Allison & Gorman, 1993; Faith et al., 1996).	<p>ALLIS-M tests for mean differences between phases after controlling for Phase A trend only. Phase A trend is semipartialled from the full dataset. A multistep procedure is followed: (a) create a temporary variable containing the scores for Phase A only, (b) regress this new "AScores" variable on Trend, (c) save the predicted output, (d) subtract these predicted values from the original Scores, (e) the resulting difference or residual scores are used in the final regression formula for ALLIS-M $R^2_{Y_{det} \cdot M}$ (<i>Y_{det}</i> = detrended response variable, <i>M</i> = 0/1 phase variable). ALLIS-M results showed no Phase B deterioration.</p>
ALLIS-MT: Allison et al.'s mean+trend model (Allison & Gorman, 1993; Faith et al., 1996).	<p>ALLIS-MT tests for simultaneous mean and trend differences between phases after controlling for Phase A trend only. The procedure is the same as for ALLIS-M, except for the final regression: $R^2_{Y_{det} \cdot TM, TM}$ (<i>Y_{det}</i> = detrended response variable, <i>M</i> = 0/1 phase variable). Thirteen ALLIS-MT results showed deteriorating Phase B trend so reverted to ALLIS-M effect sizes.</p>

Secondary analyses were then conducted to answer each research question posed: summaries of effect sizes, power analyses of effect sizes, summaries of autocorrelation coefficients, and effect size intercorrelations.

Results

Comparison of R^2 effect sizes from the nine techniques was accomplished through

boxplots, depicting percentile distributions (see Figure 1). The top and bottom wands mark 90th and 10th percentiles, and the top and bottom of the interquartile range (IQR) boxes mark 75th and 25th percentiles, encasing the median. The dots beyond the upper and lower wands are individual extreme scores.

The boxplots show great variability among the nine analytic techniques in median

values, in score variability (IRQ), and in distribution shape. Median R^2 values ranged from a low of .045 for GORS to .964 for LTD. The large differences in effect size magnitude cannot be accounted for readily. Those techniques that include both mean and trend differences (BINOM, CENT-MT, ALLIS-MT, FULL-MT, LTD) generally produced larger effect sizes, but the mean-only ALLIS-M median value (.529) was larger than those of two mean-and-trend techniques: CENT-MT (.236) and BINOM (.190). Nor can the differences in effect size magnitude be explained by whether the analytic technique is regression-based or not. The two nonregression techniques produced the largest (LTD .964) and the third smallest (BINOM .190) median R^2 s. Neither is controlling trend, a good predictor of magnitude, as the seven techniques that do so produced both very small effect sizes in GORS and CENT-M, and very large values in ALLIS-M and ALLIS-MT.

Distribution variability, indicated by the IQR also varied greatly among analytic tech-

niques. GORS effect sizes varied little across the 77 datasets, with an IQR of only .10 for the middle 50% of scores. The two ALLIS techniques, on the other hand, produced IQRs five times that width. Distribution shapes also differed, with largely symmetric distributions by ALLIS-M and SM techniques, and sharply skewed distributions by GORS and LTD. LTD showed a ceiling effect, with R^2 s at the 75th percentile above .99. In contrast, GORS showed attenuation of the distribution at the bottom of the scale. Few of the effect size distributions depicted in Figure 1 were reflective of Cohen's guidelines, which specify .01 to .14 for weak to strong results. Only GORS results were low enough for similarity with Cohen's range. This was the case even though more than 66% of the data series showed pronounced effects, and were presented by their authors as representing effective interventions. Cohen's guidelines derived from large group social science research appear not to have been appropriate for these published studies.

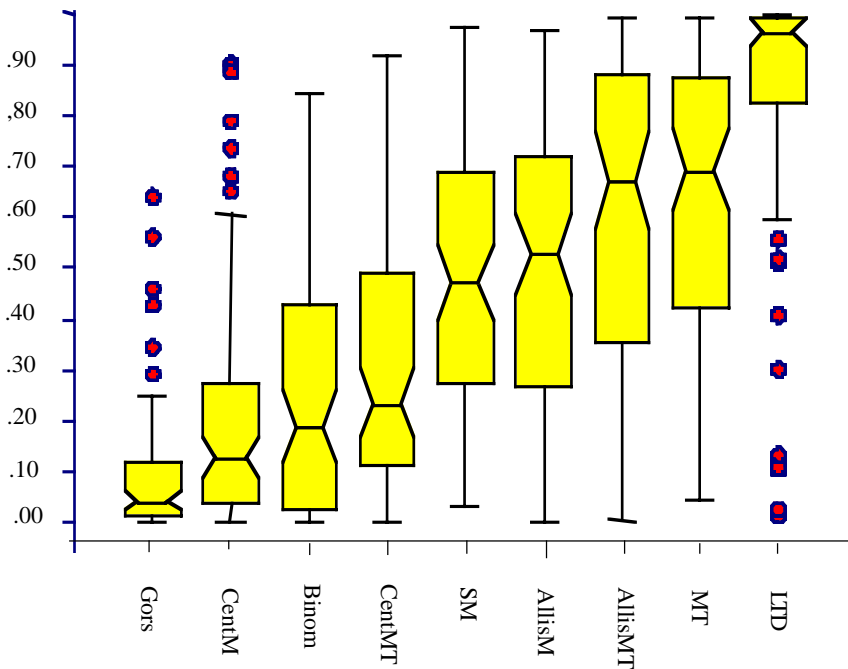


Figure 1. Percentile plots of effect sizes from nine analytic techniques applied to 77 published AB datasets.

This study also inquired about the reliability of effect sizes typical of published data, or in general terms, the power of the analytical techniques to produce statistically significant effect sizes from short data series. Power tests were conducted in the PASS module of NCSS (Hintze, 2002), which was programmed from Cohen's (1988) authoritative power analysis text. Into the PASS module are input seven values: (a) the desired power level (1- b), .80 for this study; (b) desired alpha level, .05 for this study; (c) number of predictors; (d) range of critical effect sizes of interest (which varied by analytic technique); (e) number of variables partialled out of the regression (zero or one in this study); (f) R^2 of the variables partialled out; and (g) range of number of observations of interest to the researcher (10 to 50 observations were selected). The .80 power level was selected as recommended by Cohen (1988) for most analyses. The R^2 of partialled variables refers to the trends removed in the ALLIS, CENT and GORS techniques, with most values (IQR) of $R^2 = .32$ to .76; therefore, the conservative end of that range, .32, was input into PASS. The resulting power curves

are depicted in Figure 2. Only seven of the nine analytic techniques are included, as LTD and BINOM power calculation require additional data.²

Figure 2 plots critical effect sizes (minimum effects that can be reliably detected) against the required number of observations (both A and B phases) for seven analytic techniques, in four group curves ($\beta = .2$, $\alpha = .05$). For example, for the FULL-MT technique (top line) to reliably detect effect sizes as small as $R^2 = .27$, a minimum of 30 observations would be required. The adequacy of each analytic technique for reliably detecting critical effect sizes can be evaluated by referring to the range of R^2 values from the 77 published data series. The range of the horizontal axis (number of total observations) was set 15 to 50, to include most of the sampled 77 data series, which had an IQR of 18 to 30 observations ($Mdn = 23$). It was stipulated that a useful analytic technique should be able to reliably identify at least 75% of the effect sizes encountered. Those 25th percentile R^2 values were: FULL-MT = .42, SIMP-M = .27, CENT-MT = .11, ALLIS-MT = .35, GORS = .01, CENT-M = .03, and ALLIS-M = .26.

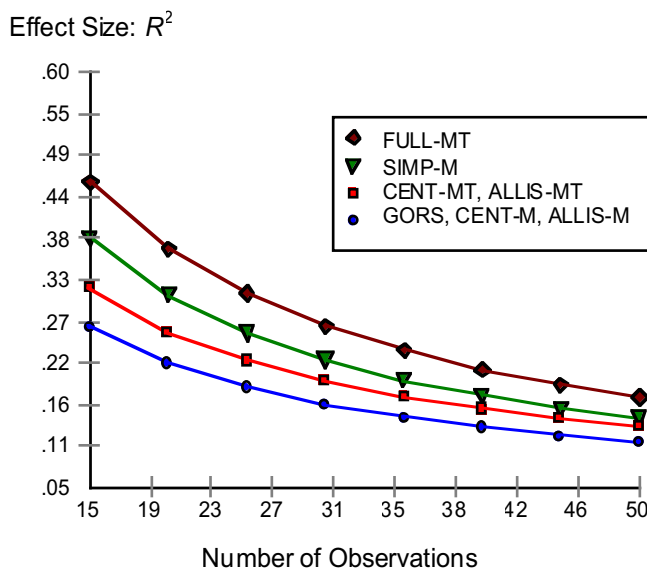


Figure 2. Power graphs ($\beta = .20$ and $\alpha = .05$) for seven single case research analysis techniques, plotted in four classes.

From the top graph line (diamonds) in Figure 2, it can be concluded that with approximately 17 observations, FULL-MT can reliably detect R^2 values as small as .42 (the 25th percentile for the obtained FULL-MT results). Because a total of 17 observations for phases A and B is considerably less than the median of 23 observations encountered, FULL-MT possesses sufficient power for most analyses. From the second curve (inverted triangles), it can be concluded that for SIMP-M to reliably detect R^2 values as small as .27 (25th percentile obtained value), approximately 25 data points are needed across phases A and B, which is close to the obtained median of 23 observations. Therefore, SIMP-M has marginally enough power for most of the published data series. The third curve (squares) shows the power of CENT-MT and ALLIS-MT techniques, both of which have two predictors and one partialled variable. This power curve shows that CENT-MT, with a 25th percentile value of .11, cannot reliably detect the targeted R^2 values in shorter data series. CENT-MT requires well over 50 data points to reliably detect R^2 values as small as .11. ALLIS-MT, on the other hand, can reliably detect its 25th percentile score (.35) with far fewer than 15 data points. The bottom power curve (circle) shows power for GORS, CENT-M, and ALLIS-M. Because of the disparate values obtained from these three techniques, conclusions for them differ. GORS requires the sensitivity to reliably detect an R^2 of .01, which it cannot approach, even with over 50 data points. CENT-M needs the power to reliably identify an R^2 of .04, which it likewise cannot do, even with 50 data points. In contrast, ALLIS-M needs to reliably detect an R^2 only as small as .27, which it can do easily with as few as 16 data points.

To summarize, of the seven techniques compared in Figure 1, four (SIMP-M, FULL-MT, ALLIS-MT, ALLIS-M) possess enough power for most of the scanned data series, with .80 power at .05 alpha level. Of these four, SIMP-M is marginally adequate. The remaining analytic techniques in Figure 2 (CENT-M, CENT-MT, GORS) lack the statistical power to reliably identify the smaller effect sizes encountered among one-fourth of the data series.

Power calculation for BINOM had to be calculated differently, based on the Z distribution for the difference between two independent proportions. Power was therefore calculated for short, average and long data series, with phase lengths 7 and 9, 10 and 11, and 14 and 16, respectively. These phase lengths represent the 25th, 50th, and 75th percentile values of the 77 data series. For short phase lengths (7 and 9), BINOM possessed only 49% power for even the most extreme phase B data split. For medium phase lengths, BINOM showed only 65% power for the most extreme phase split. For longer phase lengths, BINOM did provide adequate (80%) power, but only for the extreme phase B data splits. In summary, BINOM lacked statistical power to reliably identify effect sizes in most of the 77 published data series.

Separate power analyses were conducted for LTD, with the result that only 14 of the 77 analyses reached 85% power at the .05 significance level. Even very large effect sizes of .90 and larger were rarely statistically significant. Nor did datasets with longer phases possess higher power; phase length bore no relationship to power or statistical significance. In summary, the LTD technique possessed very low statistical power for even very large effect sizes (above .90 and .95), and longer datasets did not consistently improve power.

Problematic autocorrelation also was assayed in the seven techniques where it could be readily obtained—for all but BINOM and LTD. Percentile distributions of Lag-1 autocorrelation for the seven analytic techniques are summarized in Table 2. Following conservative cautions in the literature, autocorrelation levels greater than $r = .20$ (positive or negative) were considered to be problematic. At those levels, autocorrelation will likely distort inference (p values), and may also change effect size magnitude (Matyas & Greenwood, 1996; Parker & Brossart, 2003).

Table 2 shows that, in general, autocorrelation was a problem with all analytic techniques. Least problematic were the three techniques, which included trend: FULL-MT, ALLIS-MT, and CENT-MT. For these three, positive Lag-1 autocorrelation reached prob-

Table 2
Percentile Distributions of Residual Autocorrelations in Seven Regression Techniques from Phase A vs B Contrasts in 77 Published AB Data Series

Analytic Technique	Percentile Values				
	10 th	25 th	50 th	75 th	90 th
SIMP-M	-.217	-.023	.247	.533	.686
GORS	-.171	-.006	.230	.536	.705
ALLIS-M	-.227	-.076	.220	.542	.770
CENT-M	-.302	-.112	.123	.343	.671
ALLIS-MT	-.328	-.162	.056	.292	.641
FULL-MT	-.329	-.176	-.006	.258	.637
CENT-MT	-.333	-.188	-.008	.262	.628

lematic levels for only about 33% of the data series. In contrast, for SIMP-M, GORS, and ALLIS-M, positive autocorrelation was problematic for more than half of the analyses. Negative autocorrelation was less problematic, exhibited in 10–15% of the analyses (excluding GORS). Negative autocorrelation may be less problematic than positive, as the former is said to increase false negatives, and the latter to increase false positives in hypothesis testing (Gorman & Allison, 1996; Ostrom, 1990).

Because large differences in effect size magnitude may obscure similarities in performance among analytic techniques, intercorrelations were conducted as well. Table 3 presents a matrix of Pearson correlation coefficients, based on the 77 data series. Most analytic techniques showed moderate to high-moderate interrelationships, with tighter intercorrelated grouping or “clusters” of techniques showing coefficients of .63 to .82. Clustering of techniques was apparently not based on whether the technique was mean-based versus mean plus trend based. Instead, a relatively tight cluster (average $r = .71$) was composed of ALLIS-M, ALLIS-MT, and BINOM, with LTD more loosely joining this group. The four

members of this cluster are conceptually similar in that they all account for pre-existing Phase A trend. A second cluster of GORS, CENT-M, and CENT-MT is also noted, with an average $r = .70$. Members of this second cluster are also conceptually similar in that they all take into account the overall trend of the data series (across both phases). Finally, the remaining pair, SIMP-M and FULL-MT, neither of which controls for pre-existing trend of any kind, related at nearly $r = .80$.

Noteworthy in the matrix is the central role of the Allis techniques, especially ALLIS-MT. ALLIS-MT related most closely to the largest number of other techniques. ALLIS-MT is conceptually very similar to two very different techniques, BINOM and LTD, in that they all control for pre-existing Phase A trend. Their close interrelationships were remarkable considering their very different computational formulas.

Discussion

This study was conducted to help interpret effect sizes in single case research. The effect sizes examined were derived from nine different phase comparison analytic tech-

Table 3
Intercorrelations of Nine Statistical Techniques Performed on
62 Published AB Design Data Sets

	SIMP- M	FULL- MT	GORS	CENT- M	CENT- MT	ALLIS- M	ALLIS- MT	LTD
SIMP-M	—	.798	.520	.624	.594	.634	.490	.361
FULL-MT	.798	—	.335	.357	.505	.457	.640	.490
GORS	.520	.335	—	.680	.604	.462	.355	.175
CENT-M	.624	.357	.680	—	.817	.592	.359	.165
CENT-MT	.594	.505	.604	.817	—	.617	.538	.305
ALLIS-M	.634	.457	.462	.592	.617	—	.816	.637
ALLIS-MT	.490	.640	.355	.359	.538	.816	—	.782
LTD	.361	.490	.175	.165	.305	.637	.782	—
BINOM	.452	.586	.093	.316	.574	.614	.731	.537

niques, seven of them regression models. Single case researchers frequently publish effect sizes that do not correspond to Cohen's guidelines derived from group research, yet do not possess guidelines for the field of single case research. AB analyses were conducted on 77 published data series from several respected journals. Seventy-seven is not a large sample, considering the variability encountered in targeted behavior, phase lengths, interventions applied, and data configuration. Acknowledging the small sample, the focus of this discussion is on those findings that are more likely to be obtained in replication samples.

The first major finding from the present study was that most results did not follow Cohen's (1988) oft-cited benchmarks for "large" ($R^2 = .25$), "medium" ($R^2 = .09$), and "small" ($R^2 = .01$) effects. The GORS model yielded effects smaller than these, and the other eight models yielded much larger effects. Within the 77 published datasets, for nearly half of the analytic techniques, median R^2 values of .50 to .70 were common, with 75th percentile values of .70 to .90.

A previous study (Parker & Brossart, 2003) yielded similar findings. Those results, based on 50 fabricated "effective intervention" datasets, yielded the following median effect sizes (in parentheses). The scores without parentheses are the medians from the present study: GORS: (.028) .045; CENT-M: (.113) .130; BINOM: (.330) .190; CENT-MT: (.545) .236; ALLIS-M: (.662) .529; ALLIS-MT: (.862) .672; and LTD: (.903) .964. Across the two studies, the ranking of the techniques by median effect size is similar, providing external validation for the present study, from a very different data source.

Differences in effect size magnitude can be explained largely by their conceptual models and computational procedures. LTD, with the largest effect sizes by far, predicts differences between predicted scores in the future, on the Last Treatment Day (LTD) of phase B. Differences between predicted LTD scores were often enormous, with phase A prediction lines running nearly off the graph. However, the resulting large effect sizes possess enormous prediction error. For LTD especially, effect sizes with-

out confidence intervals appear to be relatively meaningless.

The smallest effect sizes were from GORS, CENT-M, and CENT-MT, three regression techniques similar in that they remove the full data trend prior to comparing phase performance. Faith, Allison, and Gorman (1996) note that removing full data trend is a severe adjustment, as that trend may be due to treatment in addition to pre-existing factors. The GORS trend removal is especially severe, reducing mean differences even when no trend is present. For example, a fabricated dataserie with 10 Phase A scores 2,3,2,3, etc., and 10 Phase B scores 8,9,8,9, etc., will produce these R^2 effect sizes: SIMP-M: .973, GORS: .221, and CENT-M: .893. GORS and CENT-M effect sizes are depressed by detrending, though no trends exist. The problem of overremoval of nonexistent trend does not exist with the Allis techniques, which remove only phase A trend. In summary, effect size magnitude can be accounted for partly by whether trend is removed, and, if so, by the type of trend removed.

Also determining effect size magnitude is whether trend is fully partialled or only semipartialled. This appears to account for differences between GORS and CENT-M models. GORS semipartial overall trend from client scores only (Y), whereas CENT-M fully partials trend both from both predictors (X) and scores. Semipartialing (from only the Y side of the equation) reduces the Y variance, nearly always reducing the R^2 . In the CENT techniques, the R^2 reduction is mitigated also by partialing trend from the predictors.

Another major finding of this study was that about half of the analytic techniques possessed sufficient power to detect critical effect sizes reliably. "Critical effect sizes" were defined ad hoc as those encountered in most of the 77 published data series. For seven of the techniques, a power chart could be prepared, permitting direct comparisons. BINOM and LTD required unique approaches. Four techniques, SIMP-M, FULL-MT, ALLIS-MT, and ALLIS-M, possessed sufficient power for datasets of the lengths that were scanned, although SIMP-M only marginally so. It is note-

worthy that these four techniques also produced the largest effect sizes—medians ranging $R^2 = .47 - .69$. Admonitions in the literature against using statistics with short data series held true for GORS, CENT-M, and CENT-MT, which tend to produce small effect sizes, but not for the other four regression techniques.

The separate procedures for estimating LTD and BINOM reliability yielded the same results: inadequate power for both. BINOM possessed sufficient power (80%) for reliable detection of only the most extreme results (extreme splits of the phase B data), within longer data series of approximately 30 data points and more. LTD performed worse, as even effect sizes as large as .90 and .95 could not reliably (at $\alpha = .05$) be detected with 80% power.

Regarding the problem of autocorrelation, this study confirmed what most others have found—that it does exist in most data series, often in large amounts. Defining levels larger than positive or negative .20 to be potentially problematic, over half of the results from SIMP-M, GORS, and ALLIS-M should be considered tenuous. Least autocorrelation was produced by ALLIS-MT, FULL-MT, and CENT-MT, three models with trend components. However, even for these three, more than 25% of the results were autocorrelated at potentially problematic levels. These results can be compared with those from the earlier study with fabricated data sets (Parker & Brossart, 2003). The autocorrelation values are median values from this study, and the values in parentheses are from the earlier study: GORS: (.32) .23; CENT-M: (.29) .12; ALLIS-M: (.36) .22; CENT-MT: (-.09) .008; and ALLIS-MT: (-.059) .056. The two studies show some differences, but also broad similarities. Both studies indicate that for "average" published data series, CENT-MT and ALLIS-MT are least likely to be problematic, and GORS and ALLIS-M are the most problematic. The more important conclusion, however, is that autocorrelation is a problem for most of the datasets.

In the face of undesirable levels of autocorrelation, the researcher has three alternatives: (a) drop the analysis, and rely instead on visual analysis (autocorrelation violates

even nonparametric analyses); (b) continue with the analysis, but use results only descriptively, not inferentially; or (c) cleanse the data of autocorrelation and then rerun the analysis. The second option would be more attractive were it not for the fact that moderate to high levels of autocorrelation affect not only p values, but also the R^2 sizes (Parker & Brossart, 2003). This fact makes the third option more attractive, though it is the most laborious. Though beyond the scope of this article, the present authors have successfully used ARIMA (Auto-Regressive Integrated Moving Average; Box & Jenkins, 1976; Glass, Willson, & Gottman, 1975) for short data series.

Researchers also need to know whether various analytic techniques measure very similar or different attributes. This study showed that a technique's name, purpose, or formula are not the best indicators of their resulting effect sizes. A correlation matrix provided more empirical and somewhat surprising results. Whether a technique was solely mean difference or mean plus trend difference proved important to similarity of results. An even more important determinant of covariation appeared to be whether a technique removed trend or not, whether the trend was removed from Phase A only or from the entire data series, and whether it was partialled or semipartialled. These attributes accounted for the most highly intercorrelated clusters in the matrix. In general, the matrix correlations were moderate to high-moderate in size, which gives users some assurance that similar results would be obtained even with different techniques. However, recall that these similar results were found in extreme differences in effect size magnitude.

A major finding of this study was that effect sizes need to be interpreted with respect to the technique used. Single case researchers should be informed that GORS tends to produce very small effects, and ALLIS-MT, FULL-MT, and LTD tend to produce very large effects. If effect size reliability or statistical power is a concern, then the ALLIS techniques, SIMP-M or FULL-MT may offer sufficient power with typical size data series. If statistical inference is a priority, then the researcher should also know that for techniques such as

GORS and ALLIS-M, autocorrelation tends to be problematic. Finally, some techniques may be used to replace others, as they appear to measure nearly the same thing. Inclusion of trend tended not to change the results much, as indicated by the close intercorrelations of the pairs: SIMP-M and FULL-MT, CENT-M and CENT-MT, ALLIS-M and ALLIS-MT. Of course, it is important to have an empirical or theoretical rationale for selecting a mean-only versus mean plus trend model, and for partialing or semipartialing trend or not. However, the most closely correlated techniques tabled were sometimes of different types. This study therefore provides a third, correlation-based criterion, to add to theoretical or empirical rationales for selecting a technique.

Limitations of this study are several, notably the relatively small sample of only 77 data series. Furthermore, these 77 data series represented only 26 different articles, so some dependency of results is likely among multiple graphs from a single article. Considering the small sample and its source, the agreement with findings from the previous study (Parker & Brossart, 2003) was surprising.

A second limitation to this study was the opportunistic method for selecting data series from the literature. No consideration was given to particular graph attributes (beyond the selection criteria) in choosing graphs. Instead, the first available graphs were chosen. A third limitation of this study was its restriction to AB phase comparisons. In designs of at least three phases, which are common in publication, the more interesting contrast may be something like: A_1A_2 vs. B, A_1A_2 vs. B_1B_2 , or A vs. BCD. Those contrasts were not studied, yet should be.

It can be argued that a statistical summary always adds, at minimum, an unambiguous, documentable record of effect. Yet, this study shows clearly that effect sizes themselves are anything but unambiguous—their magnitude depends largely on the technique used. In addition, some effect sizes, notably by LTD, are relatively meaningless unless constrained by information on their reliability. These cautions in the present use of statistical summaries of phase shift analyses caused the authors

of this article to ponder when statistics are and are not warranted in single case research.

Single case analysis is a relatively new arena for statistical interpretation. The standard analyses (regression, *t* tests, *Z* tests) are being used in novel ways, and with a unique class of data—short interrupted time series. The good news is the transportability of analyses and their measures of effect from group to single case research. The other good news is that some of the regression techniques appear to have sufficient power for many typical single case applications. A major challenge remaining is the problem of autocorrelation, which, though beyond the scope of this article may be remedied by using ARIMA as a backcasting, rather than forecasting tool. A second major challenge is the need for new interpretational guidelines for effect sizes. Unfortunately, a single set of benchmarks will be inadequate; several may be needed. Carefully nuanced guidelines are needed that consider the role of trend, and whether it has been partialled or semipartialled, and from which phase. Patterns are beginning to emerge, but depend for validation upon replication of studies such as this.

Footnotes

¹Omitted was the randomization design and analysis method (Edgington, 1987; Levin & Wampold, 1999), as it requires sampling across multiple phases and/or across multiple clients within a single design. Also omitted was ITSACORR (Crosbie, 1993, 1995), because its effect size results proved to be counterintuitive, bearing little relationship to results from other methods. Problems with ITSACORR have been recently documented elsewhere (Huitema, 2004).

²Power of the LTD technique depends on several variables, including the length and variability of each phase. It therefore had to be calculated for each individual dataset. Power of BINOM depends on the length of each phase in addition to the data split ratio in Phase B. It therefore was conducted for typical short, medium, and long datasets.

References

- Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behavior, Research, and Therapy*, 31, 621-631.
- Allison, D. B., Silverstein, J. M., & Gorman, B. S. (1996). Power, sample size estimation, and early stopping rules. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single case research* (pp. 335-371). Mahwah, NJ: Lawrence Erlbaum Associates.
- *Anhalt, K., McNeil, C. B., & Bahl, A. B. (1998). The ADHD classroom kit: A whole-classroom approach for managing disruptive behavior. *Psychology in the Schools*, 35, 67-79.
- Baer, D. M. (1977). Perhaps it would be better not to know everything. *Journal of Applied Behavior Analysis*, 10, 167-172.
- Barlow, D. H., & Hersen, M. (Eds.). (1984). *Single case experimental designs: Strategies for studying behavior change* (2nd ed.). Oxford, England: Pergamon Press.
- Berry, W., & Lewis-Beck, M. (Eds.). (1986). *New tools for social scientists: Advances and applications in research methods*. Beverly Hills, CA: Sage.
- Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis: Forecasting and control*. San Francisco: Holden-Day.
- *Bray, M. A. & Kehle, T. J. (1998). Self-modeling as an intervention for stuttering. *School Psychology Review*, 27, 587-598.
- *Bujold, A., Ladouceur, R., Sylvain, C., & Boisvert, J. (1994). Treatment of pathological gamblers: An experimental study. *Journal of Behavioral Therapy & Experimental Psychiatry*, 25, 275-282.
- *Burnette, M., Boehn, K., Kenyon-Jump, R., Hutton, K., & Stark, C. (1991). Control of genital herpes recurrences using progressive muscle relaxation. *Behavior Therapy*, 22, 237-247.
- Busk, P. L., & Marascuilo, L. A. (1988). Autocorrelation in single-subject research: A counterargument to the myth of no autocorrelation. *Behavioral Assessment*, 10, 229-242.
- Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single case research design and analysis: New directions for psychology and education* (pp. 187-212). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review*, 48, 378-399.
- *Cassisi, J., & McGlynn, F. (1988). Effects of EMG activated alarms on nocturnal bruxism. *Behavior Therapy*, 19, 133-142.
- Center, B. A., Skiba, R. J., & Casey, A. (1985-1986). A methodology for the quantitative synthesis of intra-subject design research. *Journal of Special Education*, 19, 387-400.
- *Chadwick, P., & Trower, P. (1996). Cognitive therapy for punishment paranoia: A single case experiment. *Behavioral Research and Therapy*, 4, 351-356.
- *Chadwick, P. D. (1994). Examining specific cognitive change in cognitive therapy for depression: A controlled case experiment. *Journal of Cognitive Psychotherapy: An International Quarterly*, 8, 19-31.
- *Chadwick, P. D., & Lowe, C. F. (1990). Measurement and modification of delusional beliefs. *Journal of Consulting and Clinical Psychology*, 58, 225-232.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

- Crosbie, J. (1987). The inability of the binomial test to control type I error with single-subject data. *Behavioral Assessment*, 9, 141-150.
- Crosbie, J. (1993). Interrupted time-series analysis with brief single-subject data. *Journal of Consulting & Clinical Psychology*, 61, 966-974.
- Crosbie, J. (1995). Interrupted time-series analysis with short series: Why it is problematic; How it can be improved. In J. M. Gottman (Ed.), *The analysis of change* (pp. 361-395). Mahwah, NJ: Lawrence Erlbaum Associates.
- Darlington, R. B., & Carlson, P. M. (1987). *Behavioral statistics: Logic & methods*. New York: Free Press.
- Edgington, E. S. (1987). Randomizing single subject experiments and statistical tests. *Journal of Counseling Psychology*, 34, 437-442.
- Faith, M. S., Allison, D. B., & Gorman, B. S. (1996). Meta-analysis of single case research. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single case research* (pp. 245-277). Mahwah, NJ: Lawrence Erlbaum Associates.
- Fidler, F., & Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed- and random-effects effect sizes. *Educational and Psychological Measurement*, 61, 575-604.
- Fowler, R. J. (1985). Point estimates and confidence intervals in measures of association. *Psychological Bulletin*, 98, 160-165.
- Fox, J. (1991). *Regression diagnostics*. Newbury Park, CA: Sage.
- Franklin, R. D., Allison, D. B., & Gorman, B. S. (Eds.). (1996). *Design and analysis of single case research*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Glass, G. V., Willson, V. L., & Gottman, J. M. (1975). *Design and analysis of time-series experiments*. Boulder: Colorado Associated University Press.
- Gorman, B. S., & Allison, D. B. (1996). Statistical alternatives for single case designs. In R. D. Franklin & D. B. Allison & B. S. Gorman (Eds.), *Design and analysis of single case research* (pp. 159-214). Mahwah, NJ: Lawrence Erlbaum Associates.
- Gorsuch, R. L. (1983). Three methods for analyzing time-series (N of 1) data. *Behavioral Assessment*, 5, 141-154.
- *Greene, B. F., Norman, K. R., Searle, M. S., Daniels, M., & Lubeck, R. C. (1995). Child abuse and neglect by parents with disabilities: A tale of two families. *Journal of Applied Behavior Analysis*, 28, 417-434.
- *Hartley, E. T., Bray, M. A., & Kehle, T. J. (1998). Self-modeling as an intervention to increase student classroom participation. *Psychology in the Schools*, 35, 363-372.
- Hartmann, D. P., Gottman, J. M., Jones, R. R., Gardner, W., Kazdin, A. E., & Vaught, R. S. (1980). Interrupted time-series analysis and its application to behavioral data. *Journal of Applied Behavior Analysis*, 13, 543-559.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego: Academic Press.
- Hintze, J. (2002). *NCSS 2002* [Computer Software]. Kaysville, UT: NCSS Statistical Software.
- Holtzman, W. H. (1963). Statistical methods for the study of change in the single case. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 199-211). Madison: University of Wisconsin Press.
- *Houten, R. V., & Retting, R. A. (2001). Increasing motorist compliance and caution at stop signs. *Journal of Applied Behavior Analysis*, 34, 185-193.
- Huitema, B. D. (1986). Statistical analysis and single-subject designs. In A. Poling & R. W. Fuqua (Eds.), *Research methods in applied behavior analysis: Issues and advances* (pp. 209-232). New York: Plenum.
- Huitema, B. E. (1985). Autocorrelation in applied behavior analysis: A myth. *Behavioral Assessment*, 7, 107-118.
- Huitema, B. E. (2004). Analysis of interrupted time-series experiments using ITSE: A critique. *Understanding Statistics*, 3, 27-46.
- 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.
- *Jensen, C. (1994). Psychosocial treatment of depression in women: Nine single subject evaluations. *Research on Social Work Practice*, 4, 267-282.
- Jones, R. R., Vaught, R. S., & Weinrott, M. (1977). Time-series analysis in operant research. *Journal of Applied Behavior Analysis*, 10, 151-166.
- Kazdin, A. (1984). Statistical analysis for single case experiments designs. In D. H. Barlow & M. Hersen (Eds.), *Single case experimental designs* (2nd ed., pp. 285-324). New York: Pergamon.
- Kazdin, A. E. (1982). *Single case research designs: Methods for clinical and applied settings*. New York: Oxford University Press.
- *Kern, L., & Bambara, L. (2002). Class-wide curricular modification to improve the behavior of students with emotional or behavioral disorders. *Behavior Disorders*, 27, 317-326.
- Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational & Psychological Measurement*, 56, 746-759.
- Kratochwill, T. R., & Brody, G. H. (1978). Single subject designs: A perspective on the controversy over employing statistical inference and implications for research and training in behavior modification. *Behavior Modification*, 2, 291-307.
- Kratochwill, T. R., & Levin, J. R. (Eds.). (1992). *Single case research design and analysis: New directions for psychology and education*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Kromrey, J. D., & Foster-Johnson, L. (1996). Determining the efficacy of intervention: The use of effect sizes for data analysis in single-subject research. *The Journal of Experimental Education*, 65, 73-93.
- *Ladouceur, R., Freeston, M., Gagnon, F., Thibodeau, N., & Dumont, J. (1993). Idiographic considerations in the behavioral treatment of obsessional thoughts. *Journal of Behavioral Therapy & Experimental Psychiatry*, 24, 301-310.
- *Lee, R., McComas, J. J., & Jawor, J. (2002). The effects of differential and lag reinforcement schedules on varied verbal responding by individuals with autism. *Journal of Applied Behavior Analysis*, 35, 391-402.
- *Lemanek, K. L., & Gresham, F. M. (1984). Social skills training with a deaf adolescent: implications for place-

- ment and programming. *School Psychology Review*, 13, 385-390.
- Levin, J. R., & Wampold, B. E. (1999). Generalized single case randomization tests: Flexible analyses for a variety of situations. *School Psychology Quarterly*, 14, 59-93.
- Linden Software. (1998) [Graph digitizing software.] Linden Software Ltd., Inglenook Cottage, Westoby Lane, Barrow Upon Humber, DN19 7DJ, UK.
- *Lopez, A., & Cole, C. L. (1999). Effects of a parent-implemented intervention on the academic readiness skills of five Puerto Rican Kindergarten students in an urban school. *School Psychology Review*, 28, 439-447.
- *Martens, B. K., Hiralall, A. S., & Bradley, T. A. (1997). A note to teacher: Improving student behavior through goal setting and feedback. *School Psychology Quarterly*, 12, 33-41.
- Matyas, T. A., & Greenwood, K. M. (1996). Serial dependency in single case time series. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single case research* (pp. 215-243). Mahwah, NJ: Lawrence Erlbaum Associates.
- Maxwell, S. E., Camp, C. J., & Arvey, R. D. (1981). Measures of strength of association: A comparative examination. *Journal of Applied Psychology*, 66, 525-534.
- Michael, J. L. (1974). Statistical inference for individual organism research: Mixed blessing or curse? *Journal of Applied Behavior Analysis*, 7, 647-653.
- Mitchell, C., & Hartmann, D. P. (1981). A cautionary note on the use of Omega squared to evaluate the effectiveness of behavioral treatments. *Behavioral Assessment*, 3, 93-100.
- Nourbakhsh, M. R., & Ottenbacher, K. J. (1994). The statistical analysis of single-subject data: A comparative examination. *Physical Therapy*, 74, 768-776.
- Ostrom, C. W., Jr. (1990). *Time series analysis: Regression techniques* (2nd ed.). Beverly Hills, CA: Sage.
- Parker, R., & Brossart, D. (2003). Evaluating single case research data: A comparison of seven statistical methods. *Behavior Therapy*, 34, 189-211.
- Parsonson, B. S., & Baer, D. M. (1992). The visual analysis of data, and current research into the stimuli controlling it. In T. R. Kratochwill & J. R. Levin (Eds.), *Single case research design and analysis* (pp. 15-40). Hillsdale, NJ: Lawrence Erlbaum Associates.
- *Pray, B., Kramer, J. J., & Lindskog, R. (1986). Assessment and treatment of tic behavior: A review and case study. *School Psychology Review*, 15, 418-429.
- *Rankin, H. (1982). Control rather than abstinence as a goal in the treatment of excessive gambling. *Behavior Research and Therapy*, 20, 185-187.
- Rosenthal, R. (1991). *Meta-analytic procedures for social research* (rev. ed. Vol. 6). Newbury Park, CA: Sage.
- Rosnow, R., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, 44, 1276-1284.
- *Salend, S. J., Whittaker, C. R., Raab, S., & Giek, K. (1991). Using a self-evaluation system as a group contingency. *Journal of School Psychology*, 29, 319-329.
- Scheffé, H. (1959). *The analysis of variance*. New York: Wiley.
- *Shapiro, E. S., Albright, T. S., & Ager, C. L. (1986). Group versus individual contingencies in modifying two disruptive adolescents' behavior. *Professional School Psychology*, 1, 105-116.
- *Sharpe, T., & Lounsbury, M. (1997). The effects of a sequential behavior analysis protocol on the teaching practices of undergraduate trainees. *School Psychology Quarterly*, 1, 105-116.
- Steiger, J. H., & Fouladi, R. T. (1992). R²: A computer program for interval estimation, power calculations, sample size estimation, and hypothesis testing in multiple regression. *Behavioral Research Methods, Instruments, and Computers*, 24, 581-582.
- Suen, H. K., & Ary, D. (1987). Autocorrelation in applied behavior analysis: Myth or reality? *Behavioral Assessment*, 9, 125-130.
- *Swanson, H. L., Kozleski, E., & Stegink, P. (1987). Disabled readers' processing of prose: Do any processes change because of intervention? *Psychology in the Schools*, 24, 378-384.
- Thompson, B. (2002). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling and Development*, 80, 64-71.
- *Tollefson, N., Tracy, D. B., Johnsen, E. P., & Chatman, J. (1986). Teaching learning disabled students goal-implementation skills. *Psychology in the Schools*, 23, 194-204.
- White, D. M., Rusch, F. R., Kazdin, A. E., & Hartmann, D. P. (1989). Applications of meta-analysis in individual subject research. *Behavioral Assessment*, 11, 281-296.
- White, O. R., & Haring, N. G. (1980). *Exceptional teaching* (2nd ed.). Columbus, OH: Merrill.
- *Winborn, L., Wacker, D. P., Richman, D. M., Asmus, J., & Geier, D. (2002). Assessment of mand selection for functional communication training packages. *Journal of Applied Behavior Analysis*, 35, 295-298.

Richard Parker received his doctorate in special education at the University of Oregon in 1990 and is an Associate Professor in the Special and Bilingual Education program at Texas A&M University. His professional interests include evaluation research, alternative assessment, and international education.

Dan Brossart received his doctorate in counseling psychology from the University of Missouri-Columbia in 1996. He is an Associate Professor in the Counseling Psychology program at Texas A&M University. His research interests include intervention research and psychotherapy processes.

Kimberly Vannest received her doctorate in 2000 from the University of Baton Rouge in Curriculum and Instruction/Special Education. She is an Assistant Professor in the Special and Bilingual Education program at Texas A&M University. Her research interests include classroom interactions, effective instruction, and emotional/behavioral disabilities.

James Long received his doctorate in counseling psychology in 2003 from Texas A&M University and is currently completing a Geropsychology Post-Doctoral Fellowship at the VA Palo Alto Health Care System.

Roman De-Alba is currently completing the internship requirements for a doctorate in school psychology from Texas A&M University in the Cypress-Fairbanks ISD, Texas.

Frank Baugh received his doctorate in counseling psychology in 2004 from Texas A&M University and is currently beginning a postdoctoral residency in organization consultation and development through the National Center for Organization Development in Cincinnati, Ohio.

Jeremy Sullivan received his doctorate in school psychology in 2003 from Texas A&M University and is currently a school psychologist for Cypress-Fairbanks ISD, Texas.