Percentage of All Non-Overlapping Data (PAND): *An Alternative to PND*

Richard I. Parker, Shanna Hagan-Burke, and **Kimberly Vannest**
Texas A&M University at College Station

Although single-case researchers are not accustomed to analyzing data statistically, standards for research and accountability from government and other funding agents are creating pressure for more objective, reliable data. In addition, "evidence-based interventions" movements in special education, clinical psychology, and school psychology imply reliable data summaries. Within special education, two heavily debated single-case research (SCR) statistical indices are "percentage of non-overlapping data" (PND) and the regression effect size, R^2. This article proposes a new index—PAND, the "percentage of all non-overlapping data"—to remedy deficiencies of both PND and R^2. PAND is closely related to the established effect size, Pearson's *Phi*, the "fourfold point correlation coefficient." The PAND/*Phi* procedure is demonstrated and applied to 75 published multiple baseline designs to answer questions about typical effect sizes, relationships with PND and R^2, statistical power, and time efficiency. Confidence intervals and *p* values for *Phi* also are demonstrated. The findings are that PAND/*Phi* and PND correlate equally well to R^2. However, only PAND/*Phi* could show adequate power for most of the multiple baseline designs sampled. The findings suggest that PAND/*Phi* may meet the requirement for a useful effect size for multiple baseline and other longer designs in SCR.

Single-case researchers traditionally have relied on visual analysis of graphs for judging intervention success. Despite the well-documented unreliability of visual judgments (Brossart, Parker, Olson, & Lakshmi, 2006; DeProspero & Cohen, 1979; Harbst, Ottenbacher, & Harris, 1991; Ottenbacher, 1990; Park, Marascuilo, & Gaylord-Ross, 1990), most published single-case research (SCR) continues to rely on visual judgments, assisted by comparisons of phase means, medians, or percentages. Visual analysis also commonly includes judging the amount of data overlap between phases, which helps capture the important concept of data dispersion or variability around a center. A recent study of 124 published SCR datasets (Parker et al., 2005) found statistical analyses in only 11%, which is comparable to the 10% rate cited in earlier studies of 10 and 25 years ago (Busk & Marascuilo, 1992; Kratochwill & Brody, 1978). Thus, in acceptance of statistical analysis, the SCR field has changed little over recent decades.

For the 75 multiple baseline designs (MBDs) sampled in this study, most authors (87%) relied solely on visual analysis, with phase percentages or means calculated, but no tests of differences between these indices. Data variability around the means and percents (e.g., standard deviations) was not presented, nor were reliabilities (standard error or confidence intervals [CIs]). The summary statistics served as nonessential additions to the primary visual analysis. Of the eight designs (11%) that were evaluated statistically, six used the student *t* test, one a Freidman two-way nonparametric test, and one a repeated measures ANOVA. Several of these authors referred to visually apparent data overlap between phases, but without quantifying the overlap.

Recent changes within the fields of education and psychology bolster the arguments for more objective and reliable results. These are the movements for evidence-based interventions, practices, or treatments in fields such as special education (Odom et al., 2005), school psychology (Kratochwill & Stoiber, 2000), and clinical psychology (Chambless & Ollendick, 2001). In school and clinical psychology, these movements are setting standards for reporting intervention efficacy, including objective and statistically reliable summaries that can be interpreted across the field and that permit comparisons between studies. The field of special education also is setting design standards, but has been slower to accept statistical summaries.

The evidence-based movements in education and psychology were accelerated (or spawned) by the federal legislation No Child Left Behind (NCLB; 2001) and the Education Sciences Reform Act (ESRA; 2002). These laws are played out in the new Institute of Education Sciences (IES; Whitehurst, 2004), which has set higher standards for funded educational research, including SCR. Corresponding policies also are reflected in statements such as the National Research Council's *Scientific Research in Education* (Shavelson & Towne, 2002).

Address: Richard I. Parker, PhD, Texas A&M University, Educational Psychology, College Station, Texas, 77843-4225
email: rparker@tamu.edu

The momentum for more objective and reliable SCR results also is accelerated by the need to have SCR studies included in meta-analyses and other non-SCR publications (Horner, Carr, Halle, McGee, Odom, & Wolery, 2005). Meta-analyses of single-case studies generally are separate from those of group research, and results from the two methodologies may not agree (Forness, 2001). This inconsistency may be because in the SCR field, "the synthesis of single-participant studies remains a controversial topic" (Forness, 2001, p. 190). The meta-analyses of special education research summarized by Forness failed to use standard effect sizes, instead employing methods that discard much of the data. Mostert (2001) evaluated all special education meta-analyses to that date against minimum standards and found most to be deficient. One missing critical piece of information was "accounting for the amount of total variance explained by the treatment effect . . . ." The higher the proportion of variance accounted for, the stronger the evidence for the efficacy of the treatment or intervention" (Mostert, p. 215). The summaries of treatment effect used by most SCR meta-analyses do not meet this minimum criterion.

The PND Versus R^2 Debate

Changes in education and psychology, especially the need to include SCR data in meta-analyses, have prompted the development of methods for calculating SCR effect sizes. In this context, a spirited debate over the best SCR effect size occurred over two decades, pitting R^2 against percent of non-overlapping data (PND; Allison & Gorman, 1993; Allison & Gorman, 1994; Scruggs, Mastropieri, & Casto, 1987; Scruggs & Mastropieri, 1998; Scruggs & Mastropieri, 2001; White, 1987). R^2, the regression effect size most frequently used in all social science research (Kirk, 1996), is championed by Allison and colleagues. It is easily converted to a second popular effect size, Cohen's (1988) *d*, the standardized mean difference (Rosenthal, 1991). For an AB design, R^2 may be translated to the following: (a) "proportion of a client's score variance explained by phase differences," (b) "reduction in uncertainty (percent increase in prediction ability) due to phase differences," (c) "the percent of the scores of one phase exceeded by the upper half of the scores of the other phase," and (d) "percent of non-overlap of client scores between phases" (Cohen, 1988, p. 22). Thus, the R^2 effect size shares some meaning with PND.

The competing index of effect was PND, championed by Scruggs, Mastropieri, and Casto (1987). PND is the percentage of Phase B data that are more extreme (in an improvement direction) than the single most extreme Phase A data point. PND can be hand-calculated from a printed graph (Scruggs & Mastropieri, 1994). Scruggs and Mastropieri (1994) offered general interpretational guidelines of PND > 70 for effective interventions, 50 < PND < 70 for interventions of questionable effectiveness, and PND < 50 for interventions with no observed effect. Gauged by acceptance in the field, PND won this debate. Of 15 single-case meta-analyses found by Scruggs and Mastropieri (2001), two thirds used PND, and only two used regression. Further supporting PND is its endorsement by recognized text authors Kazdin (1982) and Tawney and Gast (1984), as well as some meta-analysts (Kavale, Mathur, Forness, Quinn, & Rutherford, 2000).

The debate on PND versus regression highlighted their strengths and weaknesses. A summary from the debate articles indicates that PND offers at least three advantages. First is ease of calculation, as PND can be conducted with a pencil and ruler on a printed graph, and as a percentage calculation. Second is acceptability to visual analysts, as PND's emphasis on overlapping data reflects a key component of most visual analyses. The third advantage is PND's applicability to any SCR design.

The debate also helped define at least four limitations of PND, some acknowledged by its authors. First, PND is neither an effect size nor related to an accepted effect size, so it needs its own interpretation guidelines. Second, PND has unknown reliability, as it lacks a known sampling distribution, so *p* values and confidence intervals cannot be calculated. The third weakness is that PND ignores all phase A data except for one data point, which because of its extremity, is likely the most unreliable. The fourth limitation is that PND lacks sensitivity or discrimination ability, as it nears 100%, for very successful interventions.

The competing regression approach (R^2 index) advocated by Allison and Gorman (1993) demonstrates four major strengths. First, it results in R^2 or Cohen's *d* effect sizes, which are well established within the broader research community. Second, it permits calculation of confidence intervals to indicate the effect size's trustworthiness or reliability. Third, regression uses all data in both phases of SCR. Fourth, the regression approach can be expanded for more complex analyses (e.g., including phase trends; Bloom, Fischer, & Orme, 2003).

The debate also highlighted at least three limitations of the regression approach. The first limitation is that the parametric data assumptions of normality, equal variance, and serial independence are commonly not met by SCR data. Second, regression analyses can be unduly influenced by extreme outlier scores. The third limitation is that expertise is required to conduct and interpret regression analyses and to judge whether data assumptions have been met.

PAND and *Phi* or Phi^2

This article introduces an alternative index, the "percentage of all non-overlapping data" (PAND), and allied indices from the same 2 × 2 table: *Phi* and Phi^2, and examines their technical adequacy. Both *Phi* and Phi^2 are emphasized in this article, because a recent movement favors the unsquared terms (*R* and *Phi*), though the squared coefficient is still more commonly published (Abelson, 1985; Rosenthal, Rosnow, & Rubin,

2000). Like PND, PAND reflects data non-overlap between phases but differs in important respects. PAND uses all data from both phases, avoiding the criticism leveled at PND for wastefulness and for overemphasis on one unreliable data point. Importantly, PAND can be translated to Pearson's *Phi* and Phi^2, which are both "bona fide effect sizes" (Cohen, 1988, p. 223). *Phi* is a Pearson *R* for a 2 × 2 contingency table, so Phi^2 and R^2 are equivalent. Standing alone, PAND lacks status, as does PND. However, *Phi* and Phi^2 have known sampling distributions, so *p* values are available, statistical power can be estimated, and CIs can be included to indicate effect size reliability (Cohen, 1988; Fleiss, 1981). Also, Phi^2 can be transformed to Cohen's *d,* a recognized effect size in another metric.

The data requirements for PAND are minimal—just those for a chi-square test with frequency data (Cohen, 1988; Hays, 1988)—mainly, a minimum of 20 data points (5 per cell of the 2 × 2 table). The parametric requirements of equal variance and normality do not apply. The requirement of serial independence or lack of autocorrelation has little impact on PAND results because the tabled frequency data are unordered.

Unbalanced 2 × 2 tables can cause problems in calculating valid effect sizes, but the method espoused here produces tables that always have equal marginal proportions. The method of calculating overlapping data points (the minimum number that would have to be swapped across phases to eliminate all overlap) always yields an equal number in baseline and intervention phases. The equality of these two values produces a 2 × 2 table with exactly balanced marginals.

Two limitations cited for PND are not remedied by PAND. The first is insensitivity at the upper end of the scale. When there is no data overlap between Phases A and B, both PND and PAND award a 100% score, regardless of the distance between the two data clusters. PAND's second limitation is that it measures simple mean level shifts, not controlling for positive baseline trend. Like PND, it does not try to adjust for prior rate of improvement. Any positive baseline trend must be considered before attempting to infer a causal link between intervention and behavior (Parker, Cryer, & Burns, 2006). A large effect size alone does not imply that change was due to the intervention.

Although PAND avoids parametric data assumptions, its major disadvantage compared with an interval-level analysis R^2 is the reduced statistical power of a nominal-level technique. However, this reduced power will be problematic only if it prevents reliable detection of effect sizes large enough to be considered important. That is an empirical question that can be answered only from examining actual published datasets to determine the effect sizes that typically result. This study does just that, comparing statistical power of *Phi* and Phi^2 with that of *R* and R^2 for a sample of 75 published SCR datasets.

Given the total sample size required, the PAND/*Phi* procedure is not recommended for short, single-baseline designs of less than 20 or 25 data points. This article shows that samples of 60 to 80 are typical with MBDs, so they are well-suited for PAND analysis. The PAND/*Phi* procedure also will be useful for contrasts within longer, single-series designs, including multiphase designs, such as ABABA, although only MBDs are demonstrated here. The focus of this article is limited to testing a defensible effect size for one popular, strong, and complex SCR design.

PAND calculations are demonstrated both by hand and from a data spreadsheet. A touted strength of PND is that it can be calculated by hand from a graph. However, we found this procedure to be inaccurate for longer designs and more crowded graphs, and had to resort to the data spreadsheet for accurate calculation of both PND and PAND. PAND calculations are less complex than ANOVA or regression analyses, as PAND does not require interpretation of data assumptions output (equal variance, homogeneity, serial independence). For this study, the efficiency of conducting PAND was monitored with three experienced single-case researchers who had not previously used the technique.

PAND is presented as an alternative to PND for larger SCR datasets, as are typical in the special education literature. It is recommended for local use in school or clinic, for documentation and accountability, and for meta-analyses and other scholarly publications. This article applies the technique in detail to a single dataset and then field-tests it with 75 published SCR datasets. PAND is applied only to the highly regarded multiple baseline design, which has been a particularly challenging design for the field to adequately analyze. A good presentation on MBD analysis is by Busk and Serlin (1992), who recommend calculating whether data have met parametric assumptions, and depending on those results, then applying one of four alternative methods of analysis. The alternatives (hand calculations, individual *t* tests, Wilcoxon, ANOVA) produced very different effect size magnitudes of Cohen's d = 3.56 to 5.98 for their sample data. To date, the MBD lacks a generally acceptable statistical summary.

The purpose of this article is to demonstrate PAND as a generally applicable analytic technique and to field-test it with a reasonable sample of published data. Field-testing PAND/*Phi* with 75 published MBDs helps answer questions that potential users of the technique would logically pose: (a) What *Phi* effect sizes are typical with published research data, and how are they distributed? (b) What is the reliability of these effect sizes? (c) How does PAND/*Phi* correlate with the two alternatives, PND and ordinary least squares (OLS) regression? (d) How much statistical power does *Phi* possess? and (e) How efficient is this new procedure?

Method

Demonstration Dataset

PAND/*Phi* is first demonstrated with a short, fabricated MBD dataset, created to facilitate replication. In this MBD, a note-

taking strategy seeks to improve the quality of homework by Adam, Bob, and Carol—three students with learning disabilities (LD). Homework is rated weekly on a 30-point scale. An MBD graph is presented in Figure 1, with horizontal mean lines superimposed on each phase. Summary statistics on the data are presented in Table 1.

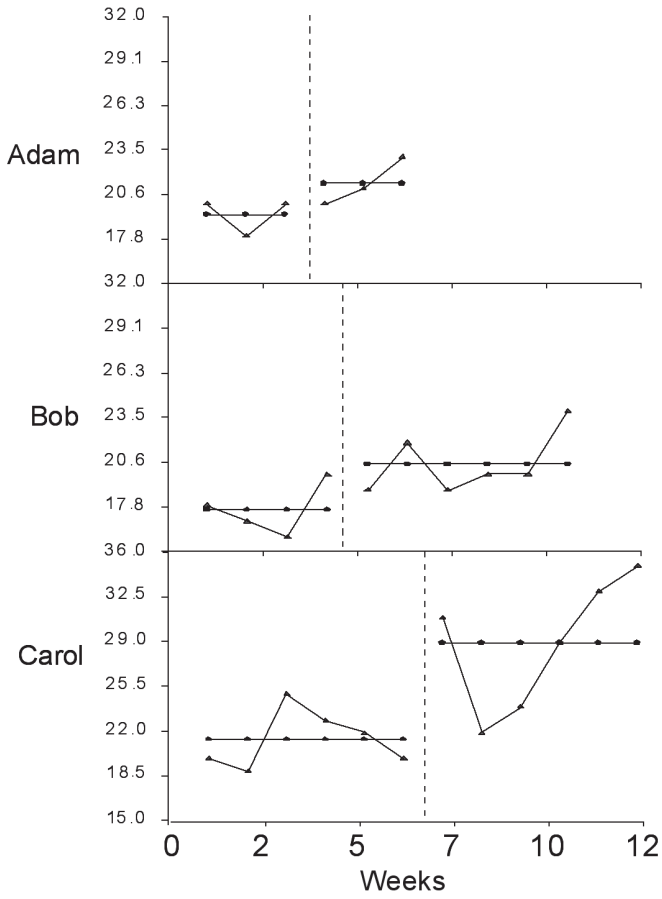

FIGURE 1. Multiple baseline across students (fabricated data).

Figure 1 and Table 1 show that the three data series for Adam, Bob, and Carol are of unequal length, as well as unequal means and variances, qualities that cause problems for an R^2 (ANOVA or regression analyses). The example was made challenging in another way: Two of the baseline phases are too short and variable to infer a stable trend. Therefore, trend control techniques available in R^2 analyses cannot be legitimately applied.

Visual analysis of Figure 1 reflects a somewhat effective intervention. Data clusters between Phase A and Phase B are not widely separated, and overlap is apparent for each student series. Overlapping data points are defined as the minimum number that would have to be swapped across phases for complete score separation. With a transparent ruler, we carefully calculated overlapping data as 2 for Adam, 2 for Bob, and 2 for Carol, totaling 6, or 6/28 = 21.4% overlap. PAND is therefore 100 – 21.4 = 78.6%. From PAND, we also can calculate non-overlap beyond chance level (50%): 78.6 – 50 = 38.6% beyond chance level.

For short datasets like this example, counting from the graph is usually accurate. For longer, more crowded datasets, an alternative spreadsheet sorting method will be provided. However, first the hand-calculation method is continued to demonstrate a 2 × 2 table and a *Phi* effect size.

Table 2 demonstrates the creation of the 2 × 2 contingency table in two steps, from left to right. Beginning at the left, the percentages of data points in the baseline and intervention phases are calculated: 13/28 = 46.4%, 15/28 = 53.6%. These percentages are entered at the bottom of their respective columns (see Table 2, left side). Next, the proportion of overlapping data (21.4%) is split between cells *b* and *c*: 10.7% in each cell. These two cells represent "too high" scores in the baseline phase (cell *b*) and "too low" scores in the intervention phase (cell *c*). Finally, cells *a* and *d* are filled in by subtraction: 46.4 – 10.7 = 35.7 and 53.6 – 10.7 = 42.9. Because this table has completely balanced vertical and horizontal marginals, a Pearson *Phi* effect size can be calculated as the difference between the two cell ratios: $[a/(a + c)] - [b/(b + d)]$. In this example 35.7/46.5 – 10.7/53.6 = .768 – .199 = .569, so *Phi* = .57. This *Phi* value can be confirmed by enter-

TABLE 1. Descriptive Summary of the Fabricated Multiple Baseline Example

	Students		
Statistics	**Adam**	**Bob**	**Carol**
Phase scores	A: 20, 18, 20 B: 20, 21, 23	A: 18, 17, 16, 20 B: 19, 22, 19, 20, 20, 24	A: 19, 18, 24, 22, 21, 19 B: 30, 21, 23, 28, 32, 34
Number of scores	6	10	12
*SD*s	1.63	2.32	5.43
Phase means	A: 19.3 B: 20.7	B: 21.3 A: 20.5	A: 17.8 B: 28.0

TABLE 2. Two-Stage Construction of the 2 × 2 Table of Proportions

	Phase A				Phase B		
Overlap	Intervention	Baseline	Total	Overlap	Intervention	Baseline	Total
Higher	*cell a*	10.7 *cell b*		Higher	35.7 *cell a*	10.7 *cell b*	46.4
Lower	10.7 *cell c*	*cell d*		Lower	10.7 *cell c*	42.9 *cell d*	53.6
Total:	46.4	53.6	100	Total:	46.4	53.6	100

ing the table's four inside numerals into either a crosstabs statistical module or a "test of two independent proportions" module; they will both yield .57. A proportions test module is preferred because nearly all provide CIs for *Phi*. For a 90% level of confidence, the exact bootstrap interval is .43 << .57 >> .71. So we can be 90% certain that the true effect size is somewhere between .43 and .71. For comparison only, a repeated measures ANOVA was conducted on the original example data. The ANOVA yielded a partial effect size for phase (A vs. B) of $R = .60$ (and $R^2 = .36$). ANOVA output indicated that our data failed to meet data assumptions of equal variance and normality, let alone the more stringent repeated measures assumption of circularity (Breckler, 1990; Keselman et al., 1998).

For longer and complex datasets, calculation of PAND from the graph may not be accurate. For those cases, an alternative spreadsheet sorting method will yield identical results. The sorting procedure is described in detail in the Appendix. As a brief overview, data are entered into four tall columns: *Random, Series, Phase*, and *Score*. *Random* is filled with random numbers. *Phase* (coded A or B) is copied and held in memory. Next, the dataset is sorted by *Random*. Then the dataset is sorted by *Score* within *Series* (a nested sort). Finally, the copied *Score* content is pasted into a new *Sorted* column. A crosstabs analysis of *Phase* and *Sorted* will yield the same results obtained earlier by hand.

Applying PAND to Published Data

PAND was applied to 75 published MBDs, obtained from 49 published articles (asterisked in *References* under the heading "Studies Sampled for Multiple Baseline Data"). In this field test, PAND was compared to PND and to R^2. Analyses were guided by research questions about the magnitude, reliability, intercorrelation, and statistical power of the new *Phi* and Phi^2 indices.

Sample Data

This study used a convenience sample of 75 multiple baseline designs. The MBD datasets were culled from a broader search within ERIC and PsycLIT from the past 20 years for all SCR designs, based on search terms such as *multiple baseline, single case, single subject, time series,* and *baseline*. The initial search revealed 104 promising MBD graphs. From those, all that were large and clear enough for digitizing were used. Only initial AB phase comparisons were analyzed in this study, although several of the designs included ABAB or ABC series. The final sample was 75 useable graphs from 49 articles.

Digitizing Graphs

The digitizing software i-extractor (Linden Software, 1998) was used to reduce published graphs back to their original data. This was accomplished in four steps. First, graphs were scanned at a resolution of 300 dots per inch (dpi) into a computer, and the resulting JPEG picture files were opened with i-extractor. Second, graph axes were set to provide actual data values on a digital Cartesian coordinate spreadsheet. Third, clicking on each data point read its value into a Microsoft Excel spreadsheet. Finally, data values were regraphed, and these graphs were compared with the originals from the articles. Reliability was checked by reprinting graphs from the digitized data; resizing them to match the originals; then overlaying the original and the recreated graphs and holding them against a bright window, which permitted quick scanning for any misplaced data points. Adjustments were required for five of the graphs, due to human error. The 75 datasets were recreated in the Number Cruncher Statistical Systems (NCSS; Hintze, 2004) statistical package. Datasets were constructed with five variable columns, for *Random, Series, ABPhase, Score,* and *Sorted*, as mentioned earlier, and described in detail in the Appendix.

Analyses

The 75 MBD graphs were analyzed by PND, PAND, and OLS regression procedures (for R^2). PND was calculated for each separate series, and the results were then averaged for a total PND score. We attempted to calculate PND directly from printed graphs with a transparent ruler, but for several graphs, the large number of crowded data points and their high vari-

ability lead to unreliable results. Therefore, PND visual analysis had to be assisted by examining *Score* and *ABPhase* columns in the data spreadsheet. PAND was calculated entirely by the sorting procedure and crosstabs analysis, as detailed in the Appendix, producing three scores: PAND, *Phi,* and Phi^2. To obtain R^2, regression analyses with the dichotomous *Phase* variable were conducted separately for each data series in a design, and the individual R^2 values were averaged together. Deteriorating series were given negative R^2 values. Regression analyses were to serve only as a ballpark external standard, without regard for meeting parametric data assumptions.

Results

Descriptive Results

Seventy-five multiple baseline designs were selected from 49 published articles. The articles are in the *References* section under the heading "Studies Sampled for Multiple Baseline Data." Each MBD possessed between two and eight separate series (or "panels" or "baselines"), with the interquartile range (IQR), or middle 50% of datasets, having three or four (Mode = 3). Each component data series averaged 23 data points, or about 11 for each of A and B phases. The average design possessed 70 data points total (IQR = 45 to 96). The smallest design had only 22 data points, and the largest had 219.

Most authors (87%) relied solely on visual analysis, some also calculating phase percentages or means, but with no tests of phase differences. When means and percentages were presented, they lacked an index of variability (SD) or reliability (Standard Error or CIs). They were nonessential adjuncts to visual analysis. Of the only eight studies (11%) with some quantitative analysis, six used the student *t* test, one used a repeated measures ANOVA, and one used a Freidman's two-way nonparametric ANOVA. None of these eight studies included effect size CIs.

Score Distributions

Figure 2 compares uniform probability (percentile) distributions of PAND (upper triangles) and PND (lower circles). The most prominent feature of this distribution graph is that both PAND and PND tend to flatten out as their values approach 100%, reflecting low discriminability. For the most successful interventions, PAND and PND distributions are quite similar, but they vary greatly for less effective interventions. The least effective interventions earned approximately PAND = 50%, which is chance-level overlap between phases, and earned PND = 0–10%. The quartile distribution for PAND was 10th: .62, 25th: .72, 50th: .84, 75th: .92, and 90th: .97. The quartile distribution for PND was 10th: .19, 25th: .23, 50th: .50, 75th: .67, and 90th: .76.

Distribution of PAND's effect size, Phi^2, is best understood in comparison with the well-known OLS R^2. Figure 3

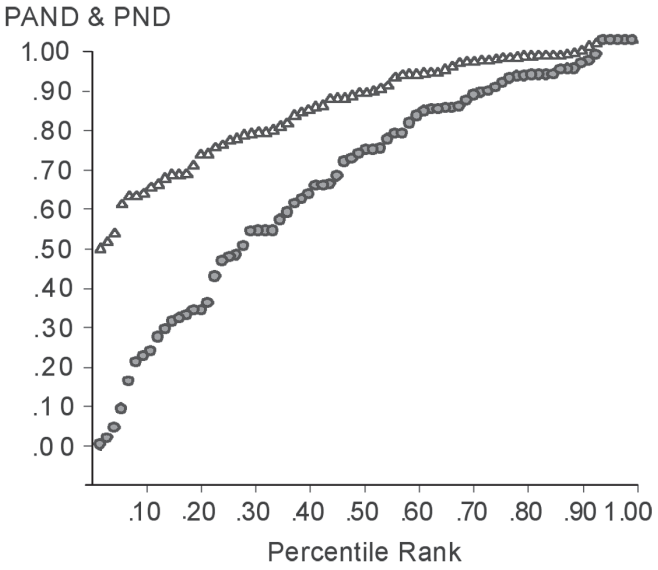

FIGURE 2. Percentile distributions of PAND (upper triangles) and PND (lower circles) for 75 published multiple baseline designs.

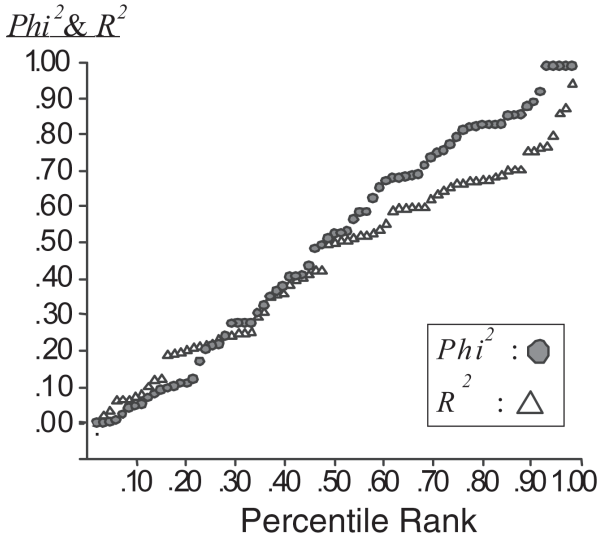

FIGURE 3. Percentile distributions of Phi^2 scores (upper circles), and R^2 (lower triangles) for 75 published multiple baseline designs.

compares their respective distributions. Phi^2 scores are the upper circles, and R^2 are the lower triangles. R^2 and Phi^2 have similar distributions until values of about .60, where Phi^2 becomes larger. The Figure 3 distribution graph reflects a ceiling effect for Phi^2 at the top end of the scale, where several scores are bunched. R^2 does not have this ceiling effect. The quartile distribution for Phi^2 was 10th: .05, 25th: .22, 50th: .53, 75th: .80, and 90th: .89. For R^2 the quartile distribution was 10th: .08, 25th: .22, 50th: .50, 75th: .67, and 90th: .76.

Thus, Phi^2 yields somewhat more extreme scores than R^2, larger for the most effective interventions, and smaller for the least effective.

Intercorrelations

The new index PAND, and its effect size Phi^2, were intercorrelated with PND and R^2 indices, resulting in Table 3. As expected, PAND and *Phi* have the highest correlation, at $R = .98$. PAND and Phi^2 also were highly correlated ($R = .95$). All other indices bore at least high-moderate relationships. The next strongest relationship (.90), between the two established effect sizes, R^2 and Phi^2, shows they measure similar attributes, though the first is continuous and the second categorical measurement. PND agreed with PAND and Phi^2 at similar levels, .85 and .84. The one lower relationship ($R = .78$), between PND and R^2, was not understandable, given that the two have little in common either in theory or in method.

Statistical Power

The relative merits of PAND and Phi^2 were next examined with regard to the criterion of statistical power (i.e., the ability to reliably detect effective interventions in typical datasets). The award-winning NCSS (Hintze, 2004) Power Analysis and Sample Size (PASS) power analysis module was used to graph power analysis curves for *Phi* and for R^2 (see Figure 4).The alpha (p level) was set at .05, and power was set at 20%. The two power analysis curves show the minimum number of MBD data points required to reliably detect a given effect size level for *Phi* (upper triangles) and for R^2 (lower circles). Figure 4 is best understood by noting $N = 69$, which was the average MBD size for our sample of 75 datasets. At $N = 69$, regression (R^2) reliably detects (at alpha = .05, and 80% power) values lower than $R^2 = .10$, reflecting strong power. Regression possesses ample power for nearly 90% of the datasets, as the 10th percentile R^2 score for the 75 datasets was .08. The *Phi* procedure, in reference to the same $N = 69$, reliably detected values as small as .34. *Phi*'s 10th percentile score was .22, and its 25th percentile score was .46. So *Phi* reliably detected more than 75% of the dataset effect sizes, but lacked the power to detect the smallest 10% to 15% of them. Thus, *Phi* had satisfactory, but not excellent, power for our sample.

In addition to statistical significance, *Phi* and Phi^2 offer a second related advantage over PND: confidence intervals. CIs are especially useful for judging the reliability of obtained effect sizes at moderate levels. The question was posed: What *Phi* CI widths are typical for published MBD data? To answer this question, exact, asymmetrical bootstrap CIs were obtained from proportions tests of five representative datasets, with *Phi* values at 90th, 75th, 50th, 25th, and 10th percentile levels. CI width depends mainly on the size of the *Phi* value and on the available sample. The data for these five datasets were all weighted to equal $N = 69$. The 95% confidence level CIs are shown bracketing obtained *Phi* scores: 90th percentile: [.79 < .94 < .99], 75th percentile: [.71 < .86 < .94], 50th percentile: [.47 < .68 < .82], 25th percentile: [.26 < .51 < .68], 10th percentile: [–.02 < .22 < .44]. Three of these five CIs are narrow enough to indicate reasonable confidence in the *Phi* score. Even down at the 25th percentile, the interval of .26 to .68 gives us some confidence in the obtained *Phi* of .51. The exception is the CI for the low 10th percentile *Phi* = .22. That CI is double the *Phi* value, and drops below zero, giving us little or no confidence in *Phi* values this low.

Efficiency

Although PND is described as a quick and easy procedure, that proved not to be the case with longer MBD datasets with high data variability. For those datasets, reliable calculations

TABLE 3. Intercorrelations Among Five Indices of Intervention Effect, Based on 75 Published Multiple Baseline Designs

Approach	R^2	PAND	*Phi*	Phi^2
R^2	1.000			
PAND	.872	1.000		
Phi	.870	.978	1.000	
Phi^2	.901	.945	.973	1.000
PND	.780	.851	.844	.836

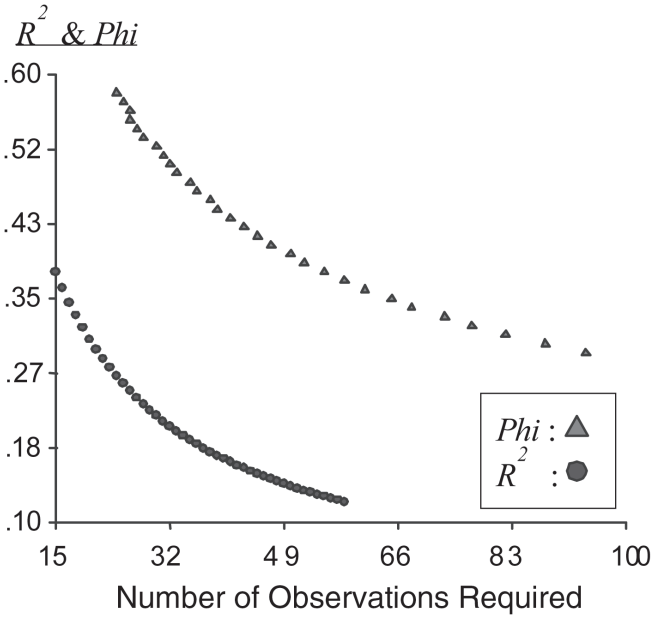

FIGURE 4. Power analysis graph for *Phi* and R^2.

from a graph with pencil and ruler were not possible, necessitating use of the data spreadsheet. So for complex designs, use of the data spreadsheet was essential for both PND and PAND, making the two procedures equal in efficiency. For shorter and uncrowded visual displays, a transparent ruler permits accurate calculation from the graph. In those cases, calculation of PND and PAND is similar in efficiency. For extra computational effort, *Phi* can be calculated along with PAND. CI calculation requires the further step of a proportions test.

Discussion

Accountability standards from funding agents and higher research and evaluation standards from government agencies are pressing researchers, including SCR practitioners, to provide objective, reliable data. Reproducible methods for summarizing SCR data also are needed for meta-analyses. The *evidence-based interventions* movements in multiple fields call for objective and reliable judgments of intervention effectiveness. Within special education, most actively debated have been two very different statistical indices: PND versus the R^2 effect size, obtained from OLS regression, *t* tests, or ANOVA. The main weaknesses of PND are that it overemphasizes a single extreme phase A data point, has no known sampling distribution, has unknown reliability, and holds little currency within the broader research community. A primary weakness of the regression approach is that single-case data too often fail to meet its parametric data assumptions.

PAND, the percentage of *all* non-overlapping data, remedies some deficiencies of both PND and R^2. PAND is a non-overlapping data index, but also is closely related, via a 2 × 2 table, to the respected Pearson's *Phi* effect size, the "fourfold point correlation coefficient." (Cohen, 1988, p. 223). *Phi* is standard output from crosstabs analyses and from tests of two independent proportions. The latter statistical modules commonly provide exact CIs for "the difference between proportions" (Cohen, 1988, p. 184), which for a balanced fourfold table is the same as *Phi*.

The PAND/*Phi* procedure was applied to a sample of 75 published multiple baseline designs. From these calculations, questions were answered regarding typical effect size magnitude, relationships to PND and R^2, statistical power, and time efficiency. One finding of this study was the high level of agreement between PND and both PAND/*Phi* indices—close to $R = .85$. However, against the OLS-based R^2, PND fared less well, at only $R = .78$, compared to .87 for the new measures. This finding is understandable, considering PND's unique procedure of focusing on a single phase A data point. PAND and *Phi* can both be calculated by hand, based on observed data overlap, but only PAND includes all data equally, reflecting more closely the non-overlap interpretation given to *R* (Cohen, 1988). Considering its unique approach to measuring data overlap, PND was a surprisingly strong performer. These results do not discount the claim of its authors that PND works well for local decision making. Scruggs and Mastropieri (1998) contended that PND agrees well with visual judgment, and that claim seems reasonable. Both PND and PAND distribution graphs showed the weakness of a ceiling effect—leveling and clumping of scores near the top of the distribution. R^2 avoids this weakness by continuing to increase as the phase scores become more widely separated, whereas PND and PAND cannot increase beyond 100% non-overlap.

Statistical power cannot be computed for PND, but it was for *Phi*, with promising results. For short, single-series datasets, *Phi* may lack power, but most of the multiple baseline designs sampled had 45 to 96 data points (IQR), reasonably balanced across phases. With these datasets, and with effect sizes of moderate magnitude, PAND/*Phi* possessed sufficient power for .05 alpha inferences. Of course, matched against R^2, PAND/*Phi*'s power is less impressive. However, significance tests and CIs are seldom needed for the weakest effect sizes.

From calculations over the 75 datasets, PAND and PND proved approximately equal in efficiency. For shorter datasets, both could be calculated from a graph with the assistance of a transparent ruler. For more complex and longer datasets and for crowded graphs, both procedures require visual scrutiny of the source data columns. Calculating confidence intervals for *Phi* requires an additional step. The additional CI information is highly desirable for publications, meta-analyses, and federal research grant-proposal writing. It is less needed for local decision making.

The R^2, Hedge's *g*, and Cohen's *d* OLS effect sizes—calculated from regression, *t* tests, or ANOVA—continue to be the most powerful analyses for multiple baseline designs. The major limitation to all three is that SCR data often fail to meet required parametric data assumptions. Failure to meet data assumptions can sometimes be addressed by data transformations or resampling/bootstrapping methodologies (Davison & Hinkley, 1997; Good, 2001). The second type of solution is to turn to nonparametric techniques, such as the categorical *Phi* summary, which entail few data assumptions. For visual analysts, PAND/*Phi* has the advantage of reflecting non-overlapping data between baseline and intervention phases. Requiring similar effort to the competing PND, PAND offers what PND cannot: (a) acceptance within the broader research community and (b) *p* values and confidence intervals to indicate reliability.

REFERENCES

Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin, 97,* 129–133.

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behaviour Research and Therapy, 31*(6), 621–631.

Allison, D. B., & Gorman, B. S. (1994). Make things as simple as possible, but no simpler: A rejoinder to Scruggs and Mastropieri. *Behaviour Research and Therapy, 32*(8), 885–890.

Bloom, M., Fischer, J., & Orme, J. G. (2003). *Evaluating practice: Guidelines for accountable professionals* (4th ed.). Needham Heights, MA: Allyn & Bacon.

Breckler, S. J. (1990). Application of covariance structure modeling in psychology: Cause for concern? *Psychological Bulletin, 107,* 260–273.

Brossart, D. F., Parker, R. I., Olson, E. A., & Lakshmi, M. (2005). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification.*

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. Levin, R. (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 159–185). Hillsdale, NJ: 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, NJ: Erlbaum.

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

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

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their applications.* Cambridge, UK: Cambridge University Press.

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

Education Sciences Reform Act of 2002 (P.L. 107–279). Retrieved September 3, 2003, from http://www.ed.gov/news/pressreleases/2002/11/11062002a .html

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: Wiley.

Forness, S. R. (2001). Special education and related services: What have we learned from meta-analysis? *Exceptionality, 9*(4), 185–197.

Good, P. I. (2001). *Resampling methods: A practical guide to data analysis.* Boston: Birkhauser.

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

Hays, W. L. (1988). *Statistics* (4th ed.). Philadelphia: Holt, Rinehart & Winston.

Hintze, J. (2004). *NCSS and PASS: Number cruncher statistical systems* [Computer software]. Kaysville, UT.

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

Kavale, K. A., Mathur, S. R., Forness, S. R., Quinn, M. M., & Rutherford, R. B., Jr. (2000). Right reason in the integration of group and single-subject research in behavioral disorders. *Behavioral Disorders, 25,* 142–157.

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

Keselman, H. J., Huberty, C., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68,* 350–386.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and 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., & Stoiber, K. C. (2000). Empirically supported interventions and school psychology: Conceptual and practical issues: Part II. *School Psychology Quarterly, 15,* 233–253.

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

Linden Software, Ltd. (1998). *i-extractor* [Graph digitizing software]. United Kingdom: Author.

Mostert, M. P. (2001). Characteristics of meta-analyses reported in mental retardation, learning disabilities, and emotional and behavioral disorders. *Exceptionality, 9*(4), 199–225.

No Child Left Behind Act of 2001, 20 U.S.C. 70 § 6301 *et seq*. (2002)

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

Ottenbacher, K. J. (1990). Visual inspection of single-subject data: An empirical analysis. *Mental Retardation, 28,* 283–290.

Park, H., Marascuilo, L., & Gaylord-Ross, R. (1990). Visual inspection and statistical analysis of single-case designs. *Journal of Experimental Education, 58,* 311–320.

Parker, R. I., Brossart, D. F., Callicott, K. J., Long, J. R., Garcia de Alba, R., Baugh, F. G., et al. (2005). Effect sizes in single case research: How large is large? *School Psychology Review, 34*(1), 116–132.

Parker, R. I., Cryer, J., & Byrns, G. (in press). Controlling trend in single case research. *School Psychology Quarterly.*

Parker, R. I., Cryer, J. & Byrns, G. (2006). Controlling trend in single case research. *School Psychology Quarterly*. 21(3) 418-440.

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: Erlbaum.

Rosenthal, R. (1991). *Meta-analysis procedures for social science research.* Beverly Hills, CA: Sage.

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

Scruggs, T. E., & Mastropieri, M. A. (1994). The utility of the PND statistic: A reply to Allison and Gorman. *Behaviour Research and Therapy, 32,* 879–883.

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

Scruggs, T. E., & Mastropieri, M. A. (2001). How to summarize single-participant research: Ideas and applications. *Exceptionality, 9,* 227–244.

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single subject research: Methodology and validation. *Remedial and Special Education, 8*(2), 24–33.

Shavelson, R. J., Towne, L., (Eds.) & the Committee on Scientific Principles for Education Research. (2002). *Scientific research in education*. Washington, DC: National Academy Press.

StatsDirect, Ltd. (2005). StatsDirect [Computer software]. Cheshire, UK: Author.

Steiger, J. H. (2005) NDC (Non-Central Distribution Calculator). Downloadable from http://www.statpower.net/page5.html

Steiger, J. H., & Fouladi, R. T. (1992). R^2: A computer program for interval estimation, power calculation, and hypothesis testing for the squared multiple correlation. *Behavior Research Methods, Instruments, & Computers, 4,* 581–582.

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

Thompson, B., Diamond, K. E., McWilliam, R., Snyder, P., & Snyder, S. W. (2005). Evaluating the quality of evidence from correlational research for evidence-based practice. *Exceptional Children, 71,* 181–194.

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

Whitehurst, G. (2004, April). *Wisdom of the head, not the heart.* Distinguished Public Policy Lecture, Institute for Policy Research, Northwestern University, Evanston, IL.

STUDIES SAMPLED FOR MULTIPLE BASELINE DATA

*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*(1), 67–79.

*Bebko, J. M., & Lennox, C. (1988). Teaching the control of diurnal bruxism to two children with autism using a simple cueing procedure. *Behavior Therapy, 19,* 249–255.

*Bray, M. A., & Kehle, T. J. (1996). Self-modeling as an intervention for stuttering. *School Psychology Review, 25,* 358–369.

*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. M. (1994). Treatment of pathological gamblers: An experimental study. *Journal of Behavioral Therapy & Experimental Psychology, 25*(4), 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.

*Chadwick, P. & Trower, P. (1996). Cognitive therapy for punishment paranoia: A single case experiment. *Behaviour Research and Therapy,* 34, 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.

*De Martini-Scully, D., Bray, M. A., & Kehle, T. J. (2000). A packaged intervention to reduce disruptive behaviors in general education students. *Psychology in the Schools, 37*(2), 149–156.

*Fantuzzo, J. W., Polite, K., & Grayson, N. (1990). An evaluation of reciprocal peer tutoring across elementary school settings. *Journal of School Psychology, 28,* 309–323.

*Gronna, S. S., Serna, L. A., Kennedy, C. H., & Prater, M. A. (1999). Promoting generalized social interactions using puppets and script training in an integrated preschool: A single-case study using multiple baseline design. *Behavior Modification, 23*(3), 419–440.

*Harris, K. R., Graham, S., Reid, R., McElroy, K., & Hamby, R. S. (1994). Self-monitoring of attention versus self-monitoring of performance: Replication and cross-task comparison studies. *Learning Disability Quarterly, 17,* 121–139.

*Harris, S. L., Handleman, J. S., & Alessandri, M. (1990). Teaching youths with autism to offer assistance. *Journal of Applied Behavior Analysis, 23,* 297–305.

*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*(4), 363–372.

*Heward, W. L., Heron, T. E., Gardner, R., & Prayzer, R. (1991). Two strategies for improving students' writing skills. In G. Stoner, M. R. Shinn, & H. M. Walker (Eds.), *Interventions for achievement and behavior problems* (pp. 379–398). Harrisonburg, VA: National Association of School Psychologists.

*Hoff, K. E., & DuPaul, G. J. (1998). Reducing disruptive behavior in general educaation classrooms: The use of self-management strategies. *School Psychology Review, 27,* 290–303.

*Houten, R. V., & Retting, R. A. (2001, Summer). Increasing motorist compliance and caution at stop signs. *Journal of Applied Behavior Analysis, 34,* 185–193.

*Jensen, C. (1994). Psychosocial treatment of depression in women: Nine single-subject evaluations. *Research on Social Work Practice, 4*(3), 267–282.

*Koegel, R. L., & Koegel, L. K. (1990). Extended reductions in stereotypic behavior of students with autism through a self-management treatment package. *Journal of Applied Behavior Analysis, 23,* 119–127.

*Laberge, B., Gauthier, J. G., Cote, G., Plamondon, J., & Cormier, H. J. (1993). Cognitive-behavioral therapy of panic disorder with secondary major depression: A preliminary investigation. *Journal of Consulting and Clinical Psychology, 61,* 1028–1037.

*Ladouceur, R., Boisvert, J. M., & Dumont, J. (1994). Cognitive-behavioral treatment for adolescent pathological gamblers. *Behavior Modification, 18,* 230–241.

*Ladouceur, R., Freeston, M. H., Gagnon, F., Thibodeau, N., & Dumont, J. (1993). Idiographic considerations in the behavioral treatment of obsessional thoughts. *Journal of Behavior Therapy and Experimental Psychiatry, 24,* 301–310.

*Lee, M. J., & Tingstrom, D. H. (1994). A group math intervention: The modification of cover, copy, and compare for group applications. *Psychology in the Schools, 31,* 133–145.

*Lemanek, K. L., & Gresham, F. M. (1984). Social skills training with a deaf adolescent: implications for placement and programming. *School Psychology Review, 13,* 385–390.

*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.

*Love, S. R., Matson, J. L., & West, D. (1990). Mothers as effective therapists for autistic children's phobias. *Journal of Applied Behavior Analysis, 23,* 379–385.

*Marlow, A. G., Tingstrom, D. H., Olmi, D. J., & Edwards, R. P. (1997). The effects of classroom-based time-in/time-out on compliance rates in children with speech/language disabilities. *Child & Family Behavior Therapy, 19*(2), 1–15.

*Marston, D. (1987). The effectiveness of special education: A time series analysis of reading performance in regular and special education settings. *The Journal of Special Education, 21*(4), 13–26.

*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*(1), 33–41.

*Montgomery, R. W., & Ayllon, T. (1994). Eye movement desensitization across subjects: Subjective and physiological measures of treatment efficacy. *Journal of Behavior Therapy and Experimental Psychiatry, 25,* 217–230.

*Mortenson, B. P., & Witt, J. C. (1998). The use of weekly performance feedback to increase teacher implementation of a prereferral academic intervention. *School Psychology Review, 27,* 613–627.

*Saigh, P. A. (1987). In vitro flooding of a childhood posttraumatic stress disorder. *School Psychology Review, 16,* 203–211.

*Saigh, P. A. (1992). The behavioral treatment of child and adolescent posttraumatic stress disorder. *Advanced Behavioral Research and Therapy, 14,* 247–275.

*Sharpe, T., & Lounsbery, M. (1997). The effects of a sequential behavior analysis protocol on the teaching practices of undergraduate trainees. *School Psychology Quarterly, 1,* 105–116.

*Skinner, C. H., Turco, T. L., Beatty, K. L., & Rasavage, C. (1989). Cover, copy, and compare: A method for increasing multiplication performance. *School Psychology Review, 18,* 412–420.

*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.

*Swanson, H. L., & Scarpati, S. (1984). Self-instruction training to increase academic performance of educationally handicapped children. *Child & Family Behavior Therapy, 6*(4), 23–39.

*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.

*Young, K. R., Morgan, D. P., & Peterson, T. J. (1988). Teaching conversation skills to behaviorally disordered children. *Psychology in the Schools, 25,* 164–174.

(*continues with Appendix*)

Appendix

Procedure for Calculating PAND/*Phi* from MBD

Datafile Setup

The PAND and *Phi* sorting/crosstabs analysis is best accomplished within a statistics package, but also can be done by Microsoft Excel. Five variable columns are created: *Random, Series, ABPhase, Score,* and *Sorted*. Into *Random* is pasted a set of random numbers. *Series* contains a different categorical tag for each series (e.g., *I, II, III, IV*). *ABPhase* is dichotomous, containing categorical tags for the two types of phases (*A, B*). *Scores* contains original scores from all series. *Sorted* is an empty column in the spreadsheet, where results from a nested sort are later pasted. The data are entered in a tall vertical column, with series under one another.

Procedure for Calculating PAND

1. *Copy ABPhase*. First, ensure the datafile is properly set up, with *Time* ascending (1, 2, 3, etc.), *Series* ascending (*I, II, III*), and *ABPhase* ascending (*A, B*) for each *Series*. When the file is properly set up, copy contents of *ABPhase,* and hold it in computer memory.
2. *Randomize*: Sort the entire dataset by the *Random* column.
3. *Nested Sort:* Sort *Score* within *Series*. If scores are expected to improve, then both variables are sorted normally, ascending. However, if *Scores* are expected to decrease across phases, then the nested *Score* is sorted inversely (descending).
4. Paste the *ABPhase* data being held in memory (copied in Step 1) into the empty *Sort* column.
5. Conduct a Crosstabs analysis on the *ABPhase* and *Sort* columns. Output will include the 2 × 2 table, as well as the *Phi* statistic. For confidence intervals around *Phi*, analyze the table's contents by a statistical module for testing two independent proportions.