Available online at www.sciencedirect.com

ScienceDirect

Behavior Therapy 40 (2009) 357–367

Behavior Therapy

www.elsevier.com/locate/bt

An Improved Effect Size for Single-Case Research: Nonoverlap of All Pairs

Richard I. Parker

Kimberly Vannest

Texas A & M University

Nonoverlap of All Pairs (NAP), an index of data overlap between phases in single-case research, is demonstrated and field tested with 200 published AB contrasts. NAP is a novel application of an established effect size known in various forms as Area Under the Curve (AUC), the Common Language Effect Size (CL), the Probability of Superiority (PS), the Dominance Statistic (DS), Mann-Whitney's U, and Sommers D, among others. NAP was compared with 3 other non-overlap-based indices: PND (percent of nonoverlapping data), PEM (percent of data points exceeding the median), and PAND (percent of all nonoverlapping data), as well as Pearson's R^2. Five questions were addressed about NAP: (a) typical NAP values, (b) its ability to discriminate among typical single-case research results, (c) its power and precision (confidence interval width), (d) its correlation with the established effect size index, R^2, and (e) its relationship with visual judgments. Results were positive, the new index equaling or outperforming the other overlap indices on most criteria.

Nonoverlapping data as an indicator of performance differences between phases has long been an important part of visual analysis in single-case research (SCR) (Sidman, 1960) and is included in recently proposed standards for evaluating SCR (Horner et al., 2005). The extent to which data in the baseline (A) versus intervention (B) phases do not overlap is an accepted indicator of the amount of performance change. Data overlap between phases has also been quantified. Twenty-five years ago, Scruggs and Casto (1987) defined "percent of nonoverlapping data" (PND) as the percent of phase B datapoints which exceed the single highest phase A datapoint. More recently, Parker, Hagan-Burke, and Vannest (2007) defined the "percent of all nonoverlapping data" (PAND) as the "percent of all data remaining after removing the minimum number of datapoints which would eliminate all data overlap between phases A and B." A third overlap index was also published recently, Ma's (2006) PEM, the "percentage of phase B datapoints exceeding the median of the baseline phase." The present paper describes a fourth index of nonoverlapping data, designed to remedy perceived weaknesses of these three. Briefly, those weaknesses are: (a) lack of a known underlying distribution (disallowing confidence intervals) (PND); (b) weak relationship with other established effect sizes (PEM); (c) low ability to discriminate among published studies (PEM, PND); (d) low statistical power for small *N* studies (PND, PAND, PEM); and (e) open to human error in hand calculations from graphs (PND, PAND, PEM).

Quantifying results in SCR is not always needed; an approximate visual judgment of "a lot" versus "little or none" may suffice for practitioners making in-house, low-stakes decisions (Parsonson & Baer, 1992). Quantifying data overlap is most needed when the interventionist needs to make more precise statements about the amount of improvement—for example, for (a) documenting evidence of clinical effectiveness for insurance, government, and legal entities; (b) comparing the relative effectiveness of two or more interventions; (c) providing support for a knowledge base of "evidence-based practices"; (d) including a study in meta-analyses; and (e) applying for competitive

Address correspondence to Richard I. Parker, Ph.D., Texas A & M University, Educational Psychology Department, College Station, TX 77843; e-mail: rparker@tamu.edu.

0005-7894/08/357–367$1.00/0

funding. These five instances are not unrelated and are an increasing part of the scientific practice of clinical and school psychology (Chambless & Ollendick, 2001; Kratochwill & Stoiber, 2002). For example, special educators and school psychologists are being charged by federal legislation (IDEA, 2004) with measuring student improvement within multitiered "Response to Intervention" models. Student response to Tier 2 and 3 interventions is commonly measured by progress monitoring, with results depicted as a line graph. But measuring the amount of change with known precision is best accomplished by an effect size index. The precision of the index is shown by confidence intervals around the effect size.

"Nonoverlap of All Pairs" (NAP), presented in this paper, summarizes data overlap between each phase A datapoint and each phase B datapoint, in turn. A nonoverlapping pair will have a phase B datapoint larger than its paired baseline phase A datapoint. NAP equals the number of comparison pairs showing no overlap, divided by the total number of comparisons. NAP can be calculated by hand from a SCR graph, from intermediate statistics available from the Wilcoxon Rank-Sum Test (Conover, 1999), or directly as Area Under the Curve (AUC) from a Receiver Operator Characteristics (ROC) diagnostic test module. Both hand calculation and computer methods are detailed in the Method section.

NAP was developed primarily to improve upon existing overlap-based effect sizes for SCR. But NAP also offers advantages over parametric analyses (*t*-tests, analyses of variance, ordinary least squares [OLS] regression), which are generally believed to possess greater statistical power. The main limitations in using parametric effect sizes are: (a) SCR data commonly fail to meet parametric assumptions of serial independence, normality and constant variance of residual scores (Parker, 2006); (b) parametric effect sizes are disproportionately influenced by extreme outliers, which are common in SCR (Wilcox, 1998); and (c) interpretation of R^2 as "percent of variance accounted for" or Cohen's *d* as "standardized mean difference" are alien to visual analysis procedures (May, 2004; Parsonson & Baer, 1992; Weiss & Bucuvalas, 1980).

Parker (2006) examined a convenience sample of 166 published SCR datasets and found that 51% failed the Shapiro-Wilk (Shapiro & Wilk, 1965) test of normality, 45% failed the Modified Levene (Brown & Forsythe, 1974) test of equal variance, and 67% failed a lag-1 autocorrelation test, using standards suggested for SCR (Matyas & Greenwood, 1996). Over three-fourths of the datasets failed one or more of these parametric assumptions (note: serial independence is also a nonparametric assumption). Given such datasets, Wilcox proclaimed standard OLS regression as, "one of the poorest choices researchers could make" (Wilcox, 1998, p. 311). He and others have challenged the accuracy of OLS effect sizes and their confidence intervals when calculated from small samples with atypical distributions, as are common in SCR. Nevertheless, R^2 is well-established in group research and generally believed to be more discriminating and to have superior statistical power, so it was included as an external standard in this study. R^2 is calculated here in an OLS linear regression module, with Phase dummy-coded (0/1).

NAP was developed mainly to improve upon existing SCR overlap-based effect sizes: PND, PAND, and PEM. NAP should offer five comparative advantages. First, NAP should discriminate better among results from a large group of published studies. Our earlier research indicated less than optimal discriminability by the other three nonoverlap indices (PND, PAND, PEM). An effect size showing equal discriminability along its full length would be most useful in comparing results within and across studies (Cleveland, 1985). The second NAP advantage should be less human error in calculations than the other three hand-calculated indices. On uncrowded graphs, PND, PAND and PEM are calculated with few errors, but not so on longer, more compacted graphs. PAND can be calculated objectively by a sorting routine, but that procedure may be confusing. NAP is directly output from an ROC module as the AUC percent and is calculated easily from Mann-Whitney U intermediate output.

A third advantage sought from NAP was stronger validation by R^2, the leading effect size in publication. Cohen's *d* was not calculated for this paper, as it can be directly computed from $R\left(R = \frac{d}{\sqrt{d^2+4}}\right)$ (Rosenthal, 1991; Wolf, 1986). Since NAP entails more data comparisons ($N_A \times N_B$) than other nonoverlap indices, it should relate more closely to R^2, which makes fullest use of data. The fourth anticipated advantage of NAP was stronger validation by visual judgments. The reason for that expectation was that visual analysis relies on multiple and complex judgments about the data, which should be difficult to capture with simpler indices such as PEM and PND. NAP is not a test on means or medians, but rather on location of the entire score distribution, and is not limited to a particular hypothesized distribution shape.

The fifth and final advantage expected from NAP was greater score precision, indicated by narrower confidence intervals (CIs). CI width is determined by the particular statistical model used, by the *N* used in calculations (narrower CIs for larger *N*s), and by effect size magnitude (narrower CIs for larger effect sizes). PEM relies on the binomial test, its *N* being the number of datapoints in phase B. PAND can be considered an effect size but is lacking in some desirable attributes. Two PAND-derived effect sizes from its 2×2 table contingency table are Pearson Phi and Risk Difference (RD; the difference between two proportions). Phi is obtained from a chi-square test, and RD from a similar "test of two proportions," both performed on the original 2×2 PAND table. Phi and RD are output with standard errors for calculating CIs, and RD is usually accompanied by CIs. Both Phi and RD analyses use as *N* the total datapoints in phases A plus B (N_A+N_B). The final nonoverlap index, PND, has no known sampling distribution, so CIs cannot be calculated. The new NAP index, output as AUC from an ROC module, is accompanied by CIs. In addition, CIs are calculated by specially developed software by Newson (2000, 2002) as add-on macros for Stata statistical software. Nonparametric AUC modules are found in most statistics packages, including NCSS, SPSS, Stata, StatExact, and SAS, which output the AUC score, its standard error of measurement, and CIs. Robust methods for NAP's CIs have been produced by Wilcox (1996) in Minitab and S-Plus. Methods for manual CI calculation are summarized by Delaney and Vargha (2002).

The overlap index here termed NAP has several names for slight variations and different areas of application (Grissom & Kim, 2005). To statisticians, it is best known in its most general form, *p* (X_1>X_2), or "the probability that a score drawn at random from a treatment group will exceed that of a score drawn at random from a control group." It can be quickly derived from U_L, or the larger of two U values from the nonparametric Mann-Whitney U test (also known as the Wilcoxon Rank-Sum test; Cliff, 1993). For diagnostic work in medicine and test development, it is "Area Under the Curve" (AUC). The "curve" in AUC is the receiver operator characteristic (ROC) curve, also known as the "sensitivity and specificity" curve, for detailing error types (false positives, false negatives) (Hanley & McNeil, 1982). Applied to continuous data it is the Common Language Effect Size (CL) (McGraw & Wong, 1992). CL was critiqued and extended by Vargha and Delaney (2000) to cover ordinal and discrete data, and they renamed the new index the "Measure of Stochastic Superiority." Similarly, Cliff (1993) promoted use a slight variation of AUC, renaming it the "Dominance Statistic" (d). Still other variations exist, often differing by how ties are handled and by their score ranges: AUC ranges from .5 to 1 (or 0 to 1 to include deteriorating scores), Cliff's d ranges from 0 to 1, and McGraw and Wong's CL ranges from −1 to +1. Since this overlap index with CIs is obtained most readily as AUC, that is the name utilized from here on.

AUC has been recommended as a broad replacement for Cohen's *d* (Acion, Peterson, Temple, & Arndt, 2006). It is popular in evidence-based medicine, in which researchers require statistics which are not saddled with parametric assumptions, are easily interpretable, and offer confidence intervals (D'Agostino, Campbell, & Greenhouse, 2006). "AUC works well with continuous, ordinal and binary variables, and is robust and maintains interpretability across a variety of outcome measures, distributions, and sample characteristics" (e.g., skewed distributions, unequal variances) (D'Agostino et al., p. 593). The score overlap interpretation of AUC is direct and intuitive. AUC is considered by some to be superior to median shift (or even mean shift), as these latter methods overemphasize central tendency. For ill-shaped distributions, the mean or median may poorly represent most data points, so an index like AUC, which emphasizes all data values equally, is preferred (Delaney & Vargha, 2002; Grissom & Kim, 2005).

Applied to SCR, NAP (AUC) can be defined as "the probability that a score drawn at random from a treatment phase will exceed (overlap) that of a score drawn at random from a baseline phase," with ties receiving one-half point. A simpler wording is "the percent of non-overlapping data between baseline and treatment phases." This concept of score overlap is identical to that used by visual analysts of SCR graphs and is the same as is calculated in the other overlap indices, PAND, PEM and PND. NAP's major theoretical advantage is that it is a comprehensive test of all possible sources of data overlap, i.e. all baseline versus all treatment datapoint comparisons, a total of $N_A \times N_B$ pairs. NAP is a probability score, normally ranging from .5 to 1. If datapoints from two phases cannot be differentiated, then AUC = .5; there is a fifty percent chance that scores from one group will exceed those of the other. For deteriorating performance during treatment phase, one must take the extra step of specifying in an AUC module the Control or Baseline phase as the high score. By doing so, the AUC range is extended to 0 to 1. Any score from 0 to .4999 represents deteriorating performance.

Given two samples with normal distributions and equal variance (unlikely in SCR), AUC or NAP can be estimated from Cohen's *d*: AUC = 1-.5∗(1 - *d*/3.464)^2 (Acion et al., 2006). The formula for estimating Cohen's d from NAP is: $d = 3.464*(1-\sqrt{(1\text{-NAP})/.5})$. So Cohen's (1988) estimates of small, medium and large d values (.2, .5, .8) correspond to NAP (on a .5 to 1 scale) values of .56, .63, and .70, respectively. And using the equivalence, $d = 2R/\sqrt{(1-R^2)}$ (Wolf, 1986), the three NAP values correspond to R^2 values of .01, .06, and .14, respectively. But our results will show that Cohen's guidelines do not apply to SCR.

This paper first illustrates with a fabricated dataset the calculation of NAP and the three other non-overlap indices. Next, all four indices, along with R^2, are applied to a sample of 200 phase AB contrasts from 122 SCR designs, published in 44 articles. This field test informs us about the new NAP index: (a) its typical values, (b) its ability to discriminate among typical single-case research results, (c) its power and precision (CI width), (d) its correlation with established indices of magnitude of effect, and (e) its relationship with visual judgments.

Method

A short AB graph has been fabricated to illustrate how the four overlap indices are calculated (see Figures 1A through D). Raw data for phase A are: 4, 3, 4, 3, 4, 7, 5, 2, 3, 2, and for phase B are: 5, 9, 7, 9, 7, 5, 9, 11, 11, 10, 9.

The PND method (Scruggs & Casto, 1987) locates the most positive datapoint in phase A and then calculates the percent of phase B datapoints that exceed it. For the example dataset, the most positive datapoint is the 6th (in time order) in phase A (circled in Figure 1A). Seven phase B datapoints exceed it, so PND = 7/11 = 64%.

The PEM method (Ma, 2006) first identifies the median level in phase A (represented by the arrow in Figure 1C) and then calculates the percent of phase B datapoints above that level. PEM = 11/11 = 100%.

The PAND method (Parker et al., 2007) finds the smallest number of datapoints from either phase whose removal would eliminate all data overlap between two phases. For these data, there is just one best solution; for other datasets multiple solutions may exist. Two datapoints need removal (circled in Figure 1B). PAND is calculated as all remaining datapoints (21-2 = 19) divided by the total N: PAND = 19/21 = 90%. PAND can be converted either to Phi, a bonafide effect size, or to Risk Difference (RD) (Cohen, 1988). To make the

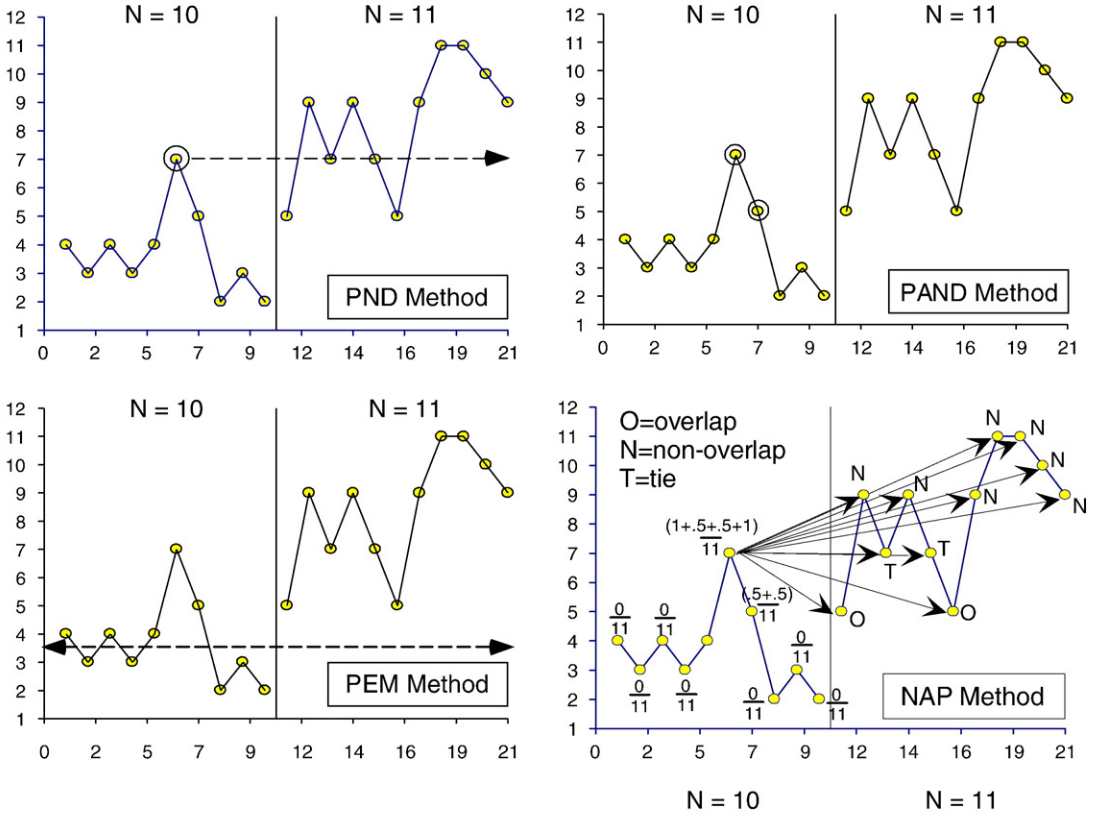

FIGURE 1 Illustration of four overlap-based methods for SCR calculating change (NAP = nonoverlap of all pairs; PND = percent of nonoverlapping data; PAND = percent of all nonoverlapping data; PEM = percent of data points exceeding the median).

conversions, datapoints needing and not needing removal from each phase are entered into a 2×2 table. Phi is obtained from a chi-square analysis of these data (Phi = .82). RD is obtained from a "two independent proportions" analysis of these data (RD = .81). In a balanced table, Phi and RD are identical. CIs are sometimes output for Phi but nearly always for RD. Calculation of Phi and RD are detailed in Parker et al. (2007), and in Parker, Vannest, and Brown (in press), respectively.

The NAP hand-calculation method (Figure 1D) compares, in turn, each phase A datapoint with each phase B datapoint. Arrows in Figure 1D show these paired comparisons and results for only one phase A datapoint, the 6th in order (value = 7). NAP hand calculation has two procedural options. One may begin counting all nonoverlapping pairs, or by counting all overlapping pairs, and subtract from the total possible pairs to obtain the nonoverlap count. The total possible pairs (total N) is the number of datapoints in phase A times phase B ($N_A \times N_B$), here 10 × 11 = 110. For most datasets, it is faster to begin counting only overlap, then subtract from the total possible pairs. The notations on Figure 1D are for counting overlap.

For the example dataset, only two of the phase A datapoints show any overlap, those with values of 7 and 5. We begin with the 7, and compare it, in turn, with each phase B datapoint. An overlap counts as one point, and a tie counts as half a point. For the Figure 1 example, comparing the 6th phase A datapoint (value = 7) with all phase B datapoints yields: (1, 0, .5, 0, .5, 1, 0, 0, 0, 0, 0) = 3. Comparing the 7th phase A datapoint (value = 5) with all phase B datapoints yields: (.5, 0, 0, 0, 0, .5, 0, 0, 0, 0, 0) = 1. Thus, the overlap sum is 3 + 1 = 4. Subtracting from the total possible pairs, we get 110 − 4 = 106. Finally, NAP = 106/110 = 96%.

NAP also is obtained directly as the AUC percent from a ROC analysis. We used NCSS (Hintze, 2006), where "ROC Curves" is located under "Diagnostic Tests." Settings to use are as follows: "Actual Condition Variable = Phase , Criterion Variable = Scores, Positive Condition Value = B (phase B), Test Direction = High × Positive." Output is "Empirical AUC = .96," along with CIs.

NAP also can be calculated from intermediate output of the Wilcoxon Rank-Sum Test, usually located in statistical packages within "Two Sample *t*-Test" or "Non-Parametric Test" modules. Wilcoxon yields a "U" value for each phase, defined as "…the number of times observations in one sample precede observations in the other sample in ranking" (Statsdirect Help, 2008). The larger U value (U_L) for phase B equals the number of nonoverlapping data comparisons (ties weighted .5). U_L divided by the total number of data comparisons ($N_A \times N_B$) equals NAP. For our example data, the Wilcoxon U_L is 106, and NAP = 106/110 = .96, the same result obtained by hand.

From our example data, similar results were obtained by the overlap indices PAND (90%), PEM (100%), and NAP (96%), whereas PND is smaller (64%). And the PAND-derived effect sizes are RD (81%) and Phi (.82). But the size of an index is less important than attributes such as its power and precision (indicated by CI width), its relatedness to established indices (criterion-related validity), its ability to discriminate among typical published results (indicated by a distribution probability plot), and its agreement with visual judgments.

CONFIDENCE INTERVALS

CIs indicate the confidence or assurance we can have in an obtained effect size, also termed "measurement precision." A wide confidence interval or band around an effect size indicates low precision, allowing little trust or assurance in that obtained value. Precision is directly related to the statistical power of a test; a test with low power cannot measure medium-sized and smaller effects with precision. CIs are strongly recommended for effect sizes by the 1999 APA Task Force on Statistical Inference (Wilkinson & The Task Force on Statistical Inference, 1999), and by the APA Publication Manual (2001). CIs are interpreted as follows: for a calculated effect size of .55, and a 90% confidence interval: .38 < .55 < .72, we can be 90% certain that the true effect size lies somewhere between .38 and .72 (Neyman, 1935). Omitted from this comparison was PND, for which CIs cannot be calculated because we do not know its chance level nor its underlying sampling distribution.

Exact CIs for PEM are provided by a binomial single-sample proportions test against a 50% chance level, easily computed from most statistics packages. From Figure 1C (11/11 = 100%), the exact 90% CI for PEM is: .87 < 1.00 < 1.00. For PAND, the same single-sample proportions test can provide CIs. The PAND of 90% is tested against a 50% chance level with $N = N_A + N_B = 21$ to yield this 90% CI: .73 < .90 < .98, a CI width of .25. But as an effect size, PAND is less suitable than two respected indices which can be calculated from a 2×2 table of PAND data: Pearson's Phi and Risk Difference (RD). From a chi-square test on the 2×2 table, Phi with 90% CI is: .46 < .82 < 1.0, a CI width of .54. From a two proportions test on the same data, RD with 90% CI is: .49 < .80 < .95, a CI width of .46. Details for calculating these Phi and RD CIs are found in Parker et al. (2007; 2009).

CIs for NAP are available with the AUC statistic as direct output. For this example, the 90% CI is: $.84 < .96 < .99$, a CI width of .15. From a total N of 18–20, we can expect reasonably accurate 90% CIs. From a total N of 30, we can expect accurate 95% CIs. And from an N of 60, accurate 99% CIs can be expected (Fahoome, 2002). Procedures have been developed for narrower CIs (greater measurement precision), which are also robust against unequal variances (Delaney & Vargha, 2002), but they are not yet available as standard output from general statistics packages. As noted earlier, some CI algorithms are available as add-ons to Stata and S-Plus.

FIELD TEST

The four overlap indices were applied to a convenience sample of 200 phase A versus B (AB) contrasts from 122 SCR designs in 44 published articles (available on request from the first author). The sample was composed of 166 AB contrasts collected 5 years ago, plus 34 contrasts recently added, to total 200. Datasets were obtained from ERIC and PsychLIT literature searches for terms such as: “single case,” “single subject,” “baseline,” “multiple baseline,” “phase,” AB, ABA, ABAB, etc. The first 300 articles located were culled to find those with graphs clear and large enough for accurate digitizing. The first 200 useable contrasts were chosen from those articles. The AB contrasts were chosen without regard for dataseries length, intervention effectiveness, design type, etc. All clear, useable graphs were digitized using I-Extractor software (Linden Software Ltd, 1998). The literature search and digitizing procedure are described in more detail in previous articles (Parker et al., 2005; Parker & Brossart, 2003).

For the 200 selected AB contrasts, the median length of a full dataseries was 18 datapoints, with an interquartile range (IQR; middle 50% of scores) of 13 to 24. Phase A had Median = 8, IQR = 5-12, and Phase B length had Median = 9, IQR = 5–13. Few articles included statistical analyses, and none provided CIs. Even for weak and moderate results, visual analysis alone was used to draw conclusions.

Results

The field test with 200 published AB contrasts was conducted to inform how NAP performs on typical datasets: (a) What are typical NAP effect size magnitudes? (b) How well does NAP discriminate among typical published SCR results? (c) How much confidence can one have in the calculated NAP values (their precision)? (d) How does NAP relate to the other indices? and (e) How well does NAP match visual judgments of client improvement? To answer these questions, NAP, PND, PAND, PEM and Pearson’s R^2 were all calculated on the 200 contrasts.

TYPICAL VALUES

Table 1 presents key percentile ranks for each of seven indices. The tabled percentile ranks show R^2 with lower scores than the four overlap indices. This is especially noticeable in the upper ranges; all four overlap indices hit their maximum (1.00) at the 90th percentile. At the 50th percentile, the overlap index values are about double R^2 , and for values at the 10th percentile, some differences are by a factor of 10. The exception is PND, which hits the floor of zero at the 10th percentile. For NAP, a full distribution would show that the 10th percentile value of .50 does not represent a floor effect. The small number of graphs with deterioration results earned scores between 0 and .50.

All of the overlap magnitudes differ enough from R^2 that they need new interpretation guidelines. For most of the studies sampled, authors described quite successful interventions, for which one would anticipate large effects. Even so, overlap results were much larger than Cohen’s guidelines for point-biserial R^2 values: "large" ($R^2 = .14$); "medium" ($R^2 = .06$); and "small" ($R^2 = .01$) (Cohen, 1988, p. 82). The tabled effect size magnitudes underscore the warning by Cohen and others that his guidelines are from large N social science group research and should not be routinely applied in other contexts (Cohen, 1988; Kirk, 1996; Maxwell, Camp, & Avery, 1981; Mitchell & Hartmann, 1981; Rosnow & Rosenthal, 1989).

DISCRIMINABILITY

The usefulness of any new index will depend in part on its ability to discriminate among results from published studies. Given a large sample, a uniform probability distribution can indicate discriminability (Cleveland, 1985). Distributions with

Table 1
Key percentile ranks for four overlap indices (NAP, PND, PAND, PEM) and the standards R^2, based on 200 published samples

	10th	Percentile Rank Values			90th
		25th	50th	75th	
NAP	.50	.69	.84	.98	1.00
PND	0.00	.24	.67	.94	1.00
PAND	.60	.69	.82	.93	1.00
PEM	.50	.79	1.00	1.00	1.00
R^2	.05	.16	.42	.65	.79

Note. NAP = nonoverlap of all pairs; PND = percent of nonoverlapping data; PAND = percent of all nonoverlapping data; PEM = percent of data points exceeding the median.

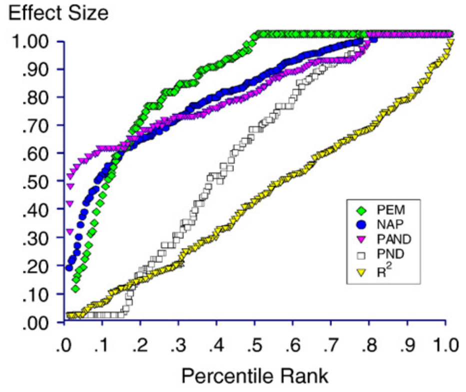

FIGURE 2 Uniform probability plots for five indices of change in SCR (NAP = nonoverlap of all pairs; PND = percent of nonoverlapping data; PAND = percent of all nonoverlapping data; PEM = percent of data points exceeding the median).

high discriminability appear as diagonal lines, without floor or ceiling effects, and without gaps, clumping, or flat segments (Chambers, Cleveland, Kleiner, & Tukey, 1983; Hintze, 2006). Figure 2 shows superior probability distributions for Pearson's R^2, forming a nearly diagonal line across the score range. The next best distributions are by NAP and PAND, both of which show ceiling effects around their 80th percentiles. So NAP and PAND do not discriminate well among the most successful 20% of interventions. PND may appear similar to these two at first glance, but it also shows a floor effect around its 18th percentile. Also, the ceiling effect of PND is more severe, near its 75th percentile. So PND discriminates poorly among 18 + 25 = 43% of the published studies. The most problematic distribution is PEM, with no floor effect, but a major ceiling effect — over 50% of the samples scored a perfect 100%.

LEVEL OF CONFIDENCE IN EFFECT SIZES

The confidence one can have in an obtained effect size is reflected in its CI width and the standard error used to calculate the CI. Narrower CIs are generally believed to come from parametric analyses such as *t*-tests, analysis of variance, and OLS regression rather than nominal-level nonparametric tests such as NAP. Ninety percent CIs were calculated for benchmark values of NAP, PAND, PEM, and R^2. CIs could not be calculated for PND because that index lacks a known sampling distribution. From the 200 datasets, CIs were calculated for small and large sample sizes ($N = 13$ and $N = 24$, respectively) and for weak and medium results (25th and 50th percentile values, respectively).

A binomial proportions test provides CIs for PEM and PAND. The PAND analysis N was $N_A + N_B$. The N for the PEM analysis was N_B only. Pearson R^2 values were bootstrap estimates from individual studies which most closely met the criteria of sample size and effect size magnitude. NAP CIs were the AUC confidence intervals from an ROC test module. Though more precise intervals are available from specialized software, they were not explored here. For R^2, exact, asymmetrical CIs were obtained from the stand-alone R^2 utility, one of the few with this capability (Steiger & Fouladi, 1992).

Table 2 contains 90% CIs for weak and medium effect sizes. For example, for weak results and a small N (13) for NAP, the Table gives: .33 <.67 <.87. This is interpreted: "We can be 90% sure that the true effect size for the obtained value of .67 is somewhere between .33 and .87 (a spread of .54 points)." PAND showed the greatest precision, its four tabled CI widths averaging .30. Second place was earned by NAP, averaging .43 over its four CIs. Ranked third was R^2, with tabled CI widths averaging .48. In last place was PEM, which produced artificially narrow CIs because two of them hit the 1.0 ceiling. Without the 1.0 ceiling, PEM's CI widths would average about .53.

RELATEDNESS TO R^2

The most commonly published effect sizes in the social sciences are members of the R^2 family (including Eta^2) (Kirk, 1996), though they entail data assumptions that single case studies often

Table 2
Ninety-percent confidence intervals on six indices of change, from small and medium sample sizes, and for weak and medium size effects

	$N = 13$		$N = 24$	
	25th %ile	50th %ile	25th %ile	50th %ile
NAP	.33 < .67 < .87	.49 < .84 < .95	.40 < .67 < .80	.62 < .84 < .94
PAND	.42 < .77 < .88	.59 < .88 < .97	.61 < .77 < .91	.70 < .88 < .96
PND	? < .26 < ?	? < .70 < ?	? < .26 < ?	? < .70 < ?
PEM	.34 < .80 < .94	.65 < 1.00 < 1.00	.53 < .80 < .93	.80 < 1.00 < 1.00
R^2	0 < .16 < .45	.04 < .42 < .68	0 < .16 < .38	.17 < .42 < .63

*Note. N*s are numbers of datapoints in a single data series, across phase A and B. NAP = nonoverlap of all pairs; PND = percent of nonoverlapping data; PAND = percent of all nonoverlapping data; PEM = percent of data points exceeding the median.

cannot meet (Parker & Brossart, 2003). Cohen's *d* is saddled with those same assumptions. In this study, R^2 served as an external standard for evaluating the new NAP. The intercorrelation matrix for the five indices is presented in Table 3. The matrix was created with Spearman's Rho rather than Pearson R to eliminate bias due to correlating a mix of squared and unsquared terms and ordinal and interval indices.

Most closely related to R^2 was NAP (Rho = .92) followed by PAND (.86), and a distance back by PEM (.75) and PND (.74). PAND and NAP were closely related (Rho = .93). Most central to the correlation matrix was NAP, which bore the overall highest relationships with most other indices. No Rho for NAP was less than .76. NAP's strong relationship with R^2 is not coincidental; an equivalence formula was presented earlier. But this formula depends on equal variances and normality of samples. So the reduction from perfect agreement to .92 reflects nonconforming sample distributions.

RELATEDNESS TO VISUAL JUDGMENTS

Three special education professors with extensive experience in SCR analysis and publication rated each of the 200 AB graphs for "amount of client improvement." There was no prior training, calibrating, or discussing among the three raters, and none had perused any of the graphs previously. A three-point scale was used for visual judgments: 1 = *little or no improvement*, 2 = *moderate improvement*, 3 = *strong or major improvement*. Agreement among the three judges ranged from r = .70 to .75, levels somewhat higher than in previous studies (DeProspero & Cohen, 1979; Harbst, Ottenbacher, & Harris, 1991; Ottenbacher, 1990; Park, Marascuilo, & Gaylord-Ross, 1990). The average ratings were correlated with each index, using Spearman's Rho to eliminate effects due to whether an index was squared or not. The resulting Rho validity coefficients were: NAP = .84, PAND = .84; R^2 = .82, PND = .71, PEM = .66. NAP and PAND distinguished themselves by correlating with visual judgments as well as R^2. PND was notably weaker, and PEM much weaker in relating to visual judgments.

Table 3
Correlation matrix for six indices of change in SCR designs

	NAP	PAND	PND	PEM
PAND	.93			
PND	.76	.73		
PEM	.81	.72	.45	
R^2	.92	.86	.74	.75

Note. NAP = nonoverlap of all pairs; PND = percent of nonoverlapping data; PAND = percent of all nonoverlapping data; PEM = percent of data points exceeding the median.

Discussion

This paper presented NAP for single-case research, a new application of the AUC nonoverlap index commonly found in medical diagnostic studies. NAP was field tested with 200 published AB phase contrasts, along with three other overlap indices (PND, PEM, PAND) and the standard, R^2. The field test addressed five questions: (a) What are typical NAP effect size magnitudes? (b) How well does NAP discriminate among typical published SCR results? (c) How much confidence can one have in the calculated NAP values (their precision)? (d) How does the NAP relate to the other indices? and (e) How well does NAP match visual judgments of client improvement?

Regarding NAP effect size magnitudes, they were large, and loosely comparable with the other overlap indices, PAND, PND, and PEM. Most NAP coefficients were at least double the R^2 values, and perfect NAP scores of 1.0 were not uncommon. NAP scores at a low 10th percentile equaled .50, and relatively few scores were below that level (reflecting deteriorating performance during intervention). NAP scores range .50 to 1.00 for nondeteriorating performance, but they can easily be transformed to a 0 to 1.00 range for nondeteriorating performance, with deterioration earning negative scores (Huberty & Lowman, 2000). PAND showed similar score magnitudes. The sharply attenuated PEM scores reached the perfect 1.0 as early as the 50th percentile rank. Given the values obtained, NAP and the other three nonoverlap indices clearly need new interpretation guidelines. Based on expert visual judgments of these 200 datasets, we can offer very tentative NAP ranges: weak effects: 0–.65; medium effects: .66–.92; large or strong effects: .93–1.0. Transforming NAP to a zero chance level gives these corresponding ranges: weak effects: 0–.31; medium effects: .32–.84; large or strong effects: .85–1.0.

On the question of the discriminability of NAP, its uniform frequency distribution was well-shaped except for a pronounced ceiling effect at the 80th percentile. None of the four overlap indices are capable of discriminating among results from the most successful intervention studies. This shortcoming of the nonoverlap indices does not exist for R^2, which can measure degree of score separation beyond complete overlap. PEM and PND distributions showed major problems which summed to an inability to discriminate among nearly half of the sample datasets. The smaller deficiencies of PAND and NAP showed them still useful for most (80%) of the sample studies, but not among the more successful interventions.

The research question about precision of results was answered in the favor of PAND, whose CI widths averaged ± .15 points, reflecting a reasonable degree of certainty. Next most satisfactory were NAP (± .21 points) and R^2 values, (± .24 points), whereas PEM was least satisfactory (approximately ± .26 points). Relative precision of the nonoverlap techniques was quite predictable from the *N* used. The surprise was the relatively wide CIs around R^2 values; however, Acion et al. (2006) document low power and precision of parametric statistics when sample data do not display constant variance and normality. Most NAP CI widths were not narrow, i.e., lacking in precision. Methods do exist for more precise NAP CIs, but they are not yet standard offerings of any statistics package, so they were not explored here.

Regarding the question of relationship to the R^2 standard, NAP was clearly superior to the other three nonoverlap indices. Besides the close relationship (Rho = .93) with R^2, NAP was the most highly related index with all others. The high .93 correlation between NAP and PAND reflects the conceptual similarity of these two indices.

The question of NAP's relationship with visual analysis judgments received a similarly positive answer. NAP tied with PAND at a substantial Rho = .84 in predicting visual judgments, a level almost equaled by R^2. The superior distribution of R^2 may be counterbalanced by the nonnormality of the data, and by the R^2 tendency to be heavily influenced by outliers.

A purpose of this study was finding an efficient and reliable calculation method. PAND has proved itself a strong index against most criteria, but hand calculation can lead to human errors, and the Excel sorting method of calculation is proving complex. NAP can be achieved without human error. It is directly output as the AUC statistic from ROC analysis with confidence intervals, and it can be calculated in one step from Mann-Whitney U output. NAP also can be calculated by hand from a graph as nonoverlapping datapoints, to give more meaning to the statistic for traditional visual analysts.

It is acknowledged that some group researchers would be hesitant to apply NAP or any analytic technique for independent groups—parametric or nonparametric—to single subject time series data. The concern is about lack of independence within the time series data. We have two main responses. First, the problem of serial dependence in SCR has been studied extensively (Hartmann et al., 1980; Huitema & McKean, 1991; Sharpley & Alavosius, 1988; Suen & Ary, 1987). It is acknowledged to exist, and it can be removed prior to analysis, although in most cases its impact on effect sizes is minor (Parker, 2006). The second response is that several respected researchers, while acknowledging the problem of serial independence, indicate those concerns are outweighed by the benefits of applying statistical analyses to phase comparisons (Matyas & Greenwood, 1996). Our conservative position is to prefer nonparametric techniques which are least affected by distribution irregularities.

Single-case research offers unique advantages for documenting progress and intervention success with atypical individuals and small groups. Establishing intervention success requires a strong research design. But documenting amount of improvement requires a strong index of magnitude of change, or effect size. Criteria for a strong effect size index for SCR include accuracy and efficiency, precision, interpretability, external validity, and valid application to ill-conforming data. The Achilles heels of Cohen's *d* and R^2 in SCR have been their unmet data assumptions and their poor interpretability. Any overlap-based index will have improved interpretability, and will require few data assumptions. Where NAP showed greatest strengths was in accuracy and efficiency of calculation and in external validation against both R^2 and visual analyst judgments. NAP did not do as well as PAND in precision (measured by CI width), which is important for small datasets. So the outcome of this study pits PAND's greater precision with NAP's greater external validity, as well as its computation efficiency and accuracy. Another advantage for NAP is that the index, variously named Area Under the Curve, Mann-Whitney $U_L/N_A \times N_B$, Common Language Effect Size, $p(X_1 > X_2)$, Probability of Superiority, Dominance Statistic, or Sommers D, etc., has a long and broad history of use. PAND is a novel index. PAND is anchored by two respected indices, Phi and Risk Difference, but they are not themselves overlap indices. PAND is closely related to these two companion indices (*R* = .84 to .92), but not identical to them.

Considering the weak showing of PND in this and earlier articles (Parker et al., 2007) and the extensive debate on its appropriateness a decade ago (Allison & Gorman, 1994), these authors question further use of that index. Similarly, the relatively new index, PEM, showed the weakest performance, confirming results of a recent PEM study (Parker & Hagan-Burke, 2007). Although PEM has now been field tested in only three studies, two of those showed performance so weak that its continued use is not recommended.

NAP's competitive performance against PAND in this study should be replicated with other published samples, but the present sample is of respectable size, among the largest for studies on single-case methods. Considering the scarcity of large-sample validations for overlap-based effect size indices, NAP could be used now. As a new, experimental index, its use should be monitored, and a cumulative record of its outcomes, strengths, and weaknesses should be reviewed regularly.

Since the greatest weakness of NAP appears to be the width of its CIs as output from AUC, other newer methods should be systematically evaluated with real data. The difficulty with Monte Carlo studies in SCR is that extreme patterns that are common in real life are rarely seen in simulations. A typical extreme yet common example from the present sample of 200 is: 0, 0, 0, 100, 10, 10, 5, 10, 100, 45, 50, 0, 0, 0, 100, 0, 0, 100. The newer analytic methods that need to be assayed include those on specialized software and those not yet integrated into software.

Limitations of this study include its restriction to simple AB contrasts and lack of consideration of data trend. The existence of prior positive trend in phase A is not uncommon in published data (Parker, Cryer & Byrns, 2006), yet it is a serious challenge to conclusion validity from simple mean or median shift tests, as well as from overlap tests. After establishing the applicability of NAP to single-case data, three major tasks are foreseen: (a) improving confidence interval estimation, (b) making NAP sensitive to linear trend in data, and (c) applying NAP to more complex designs.

References

Acion, L., Peterson, J. J., Temple, S., & Arndt, S. (2006). Probabilistic index: An intuitive non-parametric approach to measuring the size of treatment effects. *Statistics in Medicine*, *25*, 591–602.

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*, 885–890.

American Psychological Association. (2001). *Publication manual of the American Psychological Association*, 5th ed. Author: Washington, DC.

Brown, M. B., & Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, *69*, 364–367.

Chambers, J., Cleveland, W., Kleiner, B., & Tukey, P. (1983). *Graphical methods for data analysis*. Emeryville, CA: Wadsworth.

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

Cleveland, W. (1985). *Elements of graphing data*. Emeryville, CA: Wadsworth.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, *114*, 494–509.

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

Conover, W. J. (1999). *Practical nonparametric statistics*, 3rd ed. New York: Wiley.

D'Agostino, R., Campbell, M., & Greenhouse, J. (2006). The Mann–Whitney statistic: Continuous use and discovery. *Statistical Medicine*, *25*, 541–542.

Delaney, H. D., & Vargha, A. (2002). Comparing two groups by simple graphs. *Psychological Bulletin*, *79*, 110–116.

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

Fahoome, G. (2002). Twenty nonparametric statistics and their large-sample approximations. *Journal of Modern Applied Statistical Methods*, *1*(2), 248–268.

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hanley, J., & McNeil, B. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29–36.

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

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.

Hintze, J. (2006). *NCSS and PASS: Number Cruncher Statistical Systems [Computer software]*. Kaysville, UT: NCSS.

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.

Huberty, C. J., & Lowman, L. L. (2000). Group overlap as a basis for effect size. *Educational and Psychological Measurement*, *60*, 543–563.

Huitema, B. E., & McKean, J. W. (1991). Autocorrelation estimation and inference with small samples. *Psychological Bulletin*, *110*, 291–304.

Individuals with Disabilities Education Improvement Act of 2004, Pub. L. No. 108-446, 118 Stat. 2647 [Amending 20 U. S.C. § 1462 et seq.].

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746–759.

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. (1998). *I-Extractor Graph digitizing software [Computer software]*. United Kingdom: Linden Software Ltd.

Ma, H. H. (2006). An alternative method for quantitative synthesis of single-subject research: Percentage of data points exceeding the median. *Behavior Modification*, *30*, 598–617.

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., & Avery, R. D. (1981). Measures of strength of association: A comparative examination. *Journal of Applied Psychology*, *66*, 525–534.

May, H. (2004). Making statistics more meaningful for policy and research and program evaluation. *American Journal of Program Evaluation*, *25*, 525–540.

McGraw, K., & Wong, S. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*, 361–365.

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.

Newson, R. (2002). Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and median differences. *The Stata Journal*, *2*, 45–64.

Newson, R. B., 2000. snp15.1: Update to somersd. *Stata Technical Bulletin* 57: 35. In *Stata Technical Bulletin Reprints*, vol. 10, 322–323. College Station, TX: Stata Press.

Neyman, J. (1935). On the problem of confidence intervals. *Annals of Mathematical Statistics*, *6*, 111–116.

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. (2006). Increased reliability for single case research results: Is the Bootstrap the answer? *Behavior Therapy*, *37*, 326–338.

Parker, R. I., & Brossart, D. F. (2003). Evaluating single-case research data: A comparison of seven statistical methods. *Behavior Therapy*, *34*, 189–211.

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*, 116–132.

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

Parker, R. I., & Hagan-Burke, S. (2007). Median-based overlap analysis for single case data: A second study. *Behavior Modification*, *31*, 919–936.

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

Parker, R. I., Vannest, K. J., & Brown, L. (2009). The improvement rate difference for single case research. *Exceptional Children*, *75*, 135–150.

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.

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

Rosnow, R., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*, 1276–1284.

Scruggs, M., & Casto, B. (1987). The quantitative synthesis of single-subject research. *Remedial and Special Education (RASE)*, *8*, 24–33.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3 & 4), 591–611.

Sharpley, C. F., & Alavosius, M. P. (1988). Autocorrelation in behavioral data: An alternative perspective. *Behavioral Assessment*, *10*, 243–251.

Sidman, M. (1960). *Tactics of scientific research*. Boston: Authors Cooperative, Inc.

StatsDirect Ltd. Stats Direct statistical software. *http://www.statsdirect.com*. England: StatsDirect Ltd. 2008.

Steiger, J. H., & Fouladi, R. T. (1992). R^2: 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.

Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the LC common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, *25*, 101–132.

Weiss, C. H., & Bucuvalas, M. J. (1980). *Social science research and decision-making*. New York: Columbia University Press.

Wilcox, R. R. (1996). *Statistics for the social sciences*. New York: Academic Press.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist*, *53*, 300–314.

Wilkinson, L.The Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.

Wolf, F. M. (1986). *Meta-analysis: Quantitative methods for research synthesis*. Beverly Hills, CA: Sage Publications.

Received: April 6, 2008
Accepted: October 8, 2008
Available online 6 November 2008