Article

Hammill Institute on Disabilities

The Journal of Special Education
2014, Vol. 48(2) 79–91

DOI: 10.1177/0022466912456430
journalofspecialeducation
.sagepub.com

SAGE

A Simple Method to Control Positive Baseline Trend Within Data Nonoverlap

Richard I. Parker, PhD[1], Kimberly J. Vannest, PhD[1], and John L. Davis, MA[1]

Abstract

Nonoverlap is widely used as a statistical summary of data; however, these analyses rarely correct unwanted positive baseline trend. This article presents and validates the graph rotation for overlap and trend (GROT) technique, a hand calculation method for controlling positive baseline trend within an analysis of data nonoverlap. GROT is validated for controlling positive baseline trend and validated socially by visual analysis agreement. The flexibility and generality of GROT is demonstrated by using it with two alternative slope calculations: White and Haring's bi-split and Tukey's tri-split. In addition, GROT is presented as a technique that can be adapted for any non-overlap effect size method; examples here include the original percent of nonoverlapping data and newer nonoverlap of all pairs. examples here include the original percent of nonoverlapping data and newer nonoverlap of all pairs. Caution is urged to control baseline trend only when it is pronounced and reliable. GROT moves the field forward as a robust technique suitable for both visual and statistical analysis.

Keywords

single-case research, baseline trend, effect size

Judging data nonoverlap between phases is a popular technique with single-case researchers for its simplicity and close integration with the visual analysis of data graphs (Parsonson & Baer, 1978). Nonoverlap has been used for decades, and for more than 20 years, percent of nonoverlapping data (PND; Scruggs, Mastropieri, & Casto, 1987) served as a handy numeric summary of nonoverlap. A form of data nonoverlap was first used even earlier, when Owen White and Norris Haring (1980) introduced Phase B data overlap from an extended phase "celeration" or "split middle" line. The extended celeration line (ECL) for measuring growth in 6-cycle logarithmic graph paper was a key tool in precision teaching, as developed by Ogden Lindsley and his students, initially at Universities of Kansas and Washington (Calkin, 2005; Lindsley, 1991; White, 1974, 1986).

Data nonoverlap has increased legitimacy within the broader research community as it has become recognized that a complete test of nonoverlap (weighing all data points equally) has much in common with respected nonparametric "dominance statistics," Kendall's Tau, Mann–Whitney *U* test, Sommer's *d*, and area under the curve (AUC; Acion, Peterson, Temple, & Arndt, 2006; Cliff, 1993; Delaney & Vargha, 2002; Grissom & Kim, 2005). With minor calculations, output from these dominance statistics can be converted to PND (Huberty & Lowman, 2000). Thus, complete nonoverlap is now acknowledged to be a robust, distribution-free technique with good statistical power (D'Agostino, Campbell, & Greenhouse, 2006). There are currently more than nine nonoverlap techniques from which to choose (for comparisons, see Parker, Vannest, & Davis, 2011).

Although nonoverlap is easily accessible and practitioner friendly, its major shortcoming is the inability to consider baseline trend. Visual analysts warn that positive baseline trend weakens the inference that change was due to the treatment, which is termed *conclusion validity* (Kane, 2001; Kazdin, 2003; Orme, 1991). Positive baseline trend raises a competing hypothesis that progress could be due in part to preexisting improvement momentum. Because baseline trend is a serious challenge to conclusion validity, several parametric statistical methods have been designed to control for it: Crosbie's ITSACORR model (Crosbie, 1993, 1995); Last Treatment Day prediction technique of White, Rusch, Kazdin, and Hartmann (1989); mean-shift and mean-plus-trend family of models (Center, Skiba, & Casey, 1985–1986); and mean-shift and mean-plus-trend models (Allison & Gorman, 1993; Faith, Allison, & Gorman,1996).

With one notable exception, nonoverlap techniques have not attempted to control trend. The exception is ECL

[1]Texas A&M University, College Station, TX, USA

Corresponding Author:
Richard I. Parker, Texas A&M University, Department of Educational Psychology, 4225 TAMU, College Station, TX 77843-4225, USA.
E-mail: rparker@tamu.edu

introduced by White and Haring (1980). This pencil-and-ruler technique has proved itself useful since its inception nearly 40 years ago. ECL begins with hand-fitting a bi-split median line (Koenig, 1972) to Phase A data, and then extending it through Phase B. The nonoverlap calculation is the percentage of Phase B data points above that extended line, compared with an expected 50%. There are three limitations for ECL. First is the low statistical power provided by its binomial test. Second, the ECL is historically tied to Koenig's bi-split median trend line, which is presently superseded by the Tukey tri-split line (Tukey, 1977). ECL is not inherently restricted to using the bi-split line, though, and could adapt to other trend lines (though apparently adaptations have never been published). Third, ECL may not be viewed as a true nonoverlap method, as it does not directly contrast Phases A and B data points, but rather Phase B data points overlapping an extended Phase A trend line.

Despite those limitations, the ECL method of merging data nonoverlap and trend has not been equaled in nearly four decades. In fact, the recently published as percentage of data exceeding a median trend (PEM-T; Wolery, Busick, Reichow, & Barton, 2010) appears identical to ECL. Ma's percent of data exceeding the median (PEM; Ma, 2006) appears a simpler version of ECL for baseline data that have no trend (Parker & Hagan-Burke, 2007). However, ECL remains superior to PEM by offering a viable (albeit low power) analysis summary.

The field of single-case research (SCR) nonoverlap methods presently has nine contenders. They have been compared in detail elsewhere (Parker et al., 2011), so here a brief overview will suffice. The nine are as follows: (a) ECL (White & Haring, 1980), (b) PND (Scruggs et al., 1987), (c) PEM (Ma, 2006), (d) percent of all nonoverlapping data (PAND; Parker, Hagan-Burke, & Vannest, 2007), (e) Pearson's phi (φ; Parker et al., 2007), (f) improvement rate difference (IRD; Parker, Vannest, & Brown, 2009), (g) nonoverlap of all pairs (NAP; Parker & Vannest, 2009), (h) Kendall's Tau for nonoverlap between groups (TAU_{novlap}; Parker, Vannest, Davis, & Sauber, 2011), and (i) Combined Mann–Whitney *U* test nonoverlap with Tau baseline trend control (Tau-U; Parker, Vannest, et al., 2011). This list does not include percentage reduction data (PRD; O'Brien & Repp, 1990), which is a parametric and mean-based method, rather than nonoverlap. It also does not include the percentage of zero data (PZD; Scotti, Evans, Meyer, & Walker, 1991), which is applicable to only some clients and goals. The limitation of failing to consider Phase A trend applies to all but two of the nine listed, exceptions being the venerable ECL and the new Tau-U. The limitation of offering no *p* values applies only to PND. The weakness of very low statistical power (especially undesirable with small samples) applies to ECL, PND (for which statistical power is unknown), and PEM. The other six nonoverlap techniques represent improvement due to greater power, less likelihood of chance-level results, and better precision for data-based decisions (Parker et al., 2011).

The present article introduces a new visual-graphic method with the same aims as ECL, with some distinct advantages over its predecessor. The new method, graph rotation for overlap and trend (GROT), allows users to control positive baseline trend and calculate a nonoverlap-based effect size on the adjusted data set. GROT is a flexible technique that can be applied using any trend line slope estimate or nonoverlap effect size. Unlike ECL, GROT is a true nonoverlap technique, with more direct interpretability. GROT yields a "PND" summary score, more meaningful than ECL's interpretation as the "ratio of data points around an extended baseline." GROT's second advantage as a true nonoverlap method is that it offers considerably more statistical power than ECL. ECL relies on the low power binomial test or median-based "Sign Test" for proportion of data split by the extended median-based trend line. In contrast, the flexibility available in GROT allows users to apply effect size metrics with higher statistical power. For example, NAP or any other "dominance" nonoverlap test possesses 91% to 94% of the power of a regression test as measured by the "Pitman Efficiency" rating (Hollander & Wolfe, 1999). Pitman Efficiency index for the Sign Test used in the ECL method is closer to 60%, depending on sample size and distribution shape (Hodges & Lehmann, 1956). The third of GROT's advantages is its broad applicability with any trend and with any nonoverlap index. Thus, it can be adapted to user preferences, and be used with new or future trend estimates and nonoverlap indices as they are developed. Despite the differences between ECL and GROT, they have much in common. Both are graph based, relying on visual analysis and pencil-and-ruler operations on paper. Both are "distribution free" and robust to outliers. Both can be applied to ordinal as well as interval-level scales, and to data which fail to meet parametric distribution assumptions (Wilcox, 2010). Therefore, GROT provides three improvements as answers to limitations of ECL, while retaining the strengths which have promoted its longevity.

GROT is first demonstrated here on two data sets, with results technically validated by the well-reputed regression method by Allison and colleagues (Allison & Gorman, 1993; Faith et al., 1996). Next, GROT graphs are subjected to visual judgments to answer two questions. First, "Will results produced from GROT agree with visual judgment of expert but blind raters?" and second, "Will effects be reduced (between Phases A and B) due to baseline trend control?"

In brief, the GROT procedure is as follows to provide the reader a general idea of the steps. More detailed procedures appear following this brief explanation and these details correspond to some of the options for this flexible method,

including two trend calculation options and two nonoverlap options.

First, a trend line is fit to Phase A (any calculation may be used as demonstrated later). Second, it is "dropped down" to the *X*- and *Y*-axis intersect (keeping parallel with the original line), and also extended through the end point of Phase B. Third, the graph paper is rotated so that new dropped trend line becomes parallel with a new horizontal base of the graph. In addition, the line between the two phases (intervention onset line) is redrawn so that it is vertical and now perpendicular to the new horizontal axis. Finally, one can compare Phases A and B data visually and/or statistically. We later present two statistical methods as illustrations and visual judgments as validation. The physically rotated graph has the effect of statistically controlling for Phase A trend. The effect is identical to using semipartial correlation to statistically control for baseline trend, as in the well-reputed technique by Allison, Faith, and colleagues (Allison & Gorman, 1993) without the calculation.

Two Trend Line Slopes

Because GROT is a general method which works with any trend line slope estimate, this article demonstrates two rank-order slopes: the bi-split or "quarter intersect" slope (Koenig, 1972; White, 1974) and Tukey's tri-split median-based slope (Tukey, 1977). Other options for trend calculation include the Theil–Sen or "Kendall's slope" (Sen, 1968; Theil, 1950) and the linear regression slope. We do not provide examples for these but they also work with GROT.

Koenig bi-split, median-based slope. The most popular hand-fit trend line in special education was first introduced to educators by Koenig (1972), modified by White (1974), and popularized by White and Haring (1980) and by Kazdin (1982). Koenig's bi-split "quarter intersect" method was first widely used in schools within the precision teaching (Pennypacker, Koenig, & Lindsley, 1972). The slope was calculated on a "celeration line" plotted on a "standard celeration chart" (6-cycle logarithmic ruled graph paper) and is distinguished from performance rate (Calkin, 2005). Readers may be interested to note that this bi-split method of calculating celeration or slope was known even earlier outside of education, where it was called the "Brown–Mood slope," as it was popularized by Brown and Mood (1951) from an even earlier publication by Wald (1940).

The quarter intersect method entails first splitting the data vertically into earlier and later halves, and then marking the intersection of the median X and median Y values for each half. A line is then drawn to connect the two median intersects across the two phases. Optionally, the trend line may be raised or lowered (keeping parallel with the original) so it splits all data points 50% above and 50% below it.

Tukey tri-split median-based slope. The Tukey tri-split line was popularized by Tukey and colleagues from the Exploratory Data Analysis group at Princeton (Hoaglin, Mosteller, & Tukey, 1983; Tukey, 1977). The tri-split line was well known from the 1940s (Bartlett, 1949; Nair & Srivastava, 1942; Wald, 1940) and outside of education is often referred to as "Wald's trend line" or "Wald's slope" in deference to its earliest source. Tukey and colleagues, however, did the most to popularize the technique.

The Tukey (1977) tri-split slope begins with dividing the data into three equal parts on the *X*-axis, for example, data at Sessions 1 to 3, 4 to 6, and 7 to 9 on a 9-point data series: We will refer to this as early, middle, and late, respectively. The trend line is based only on the early and late thirds of the data, and the middle data portion has a limited role, adjusting the trend line up or down (keeping parallel to the original). The intersect of median X and Y values are marked for the early segment and the late data segments, and a trend line is drawn to connect the two intersects. Optionally, the line can be adjusted up or down (keeping parallel with original) so it splits all data, 50% above and below it. Another option is to raise or lower the line so it passes through the median of the middle data segment. The Tukey method is currently used to train teachers in progress monitoring, as a substitute for the bi-split method (Hintze & Stecker, 2006).

The most extensive evaluation of median-based trend techniques was a Monte Carlo study by Johnstone and Velleman (1985). The Tukey tri-split method consistently outperformed the bi-split method on power and efficiency (Pittman coefficient). In schools research, Parker, Stein, and Tindal (1992) predicted student oral reading fluency scores with bi-split, tri-split, and linear regression lines. The Tukey tri-split line was closer to the regression line and surpassed even the linear regression line in predicting actual future performance.

Another promising trend line, not included in this article, is the "Theil–Sen slope" (Sen, 1968; Theil, 1950), also known as "Kendall's robust line-fit method" (Sokal & Rohlf, 1995). Theil–Sen is the median slope of many "mini-slopes" created from all pairwise data comparisons made in time order (early to late) in a time series. It is available in an increasing number of free applications: the free student MYSTAT software (SYSTAT, 2008), the freely downloadable WinPEPI software for health care and medical researchers (Abramson, 2010), and the free software KTRLine Version 1.0 (Granato, 2006) from the U.S. Geological Survey Office. Although not yet used in schools research, it is mentioned here because of its future promise. In the Johnstone and Velleman (1985) Monte Carlo study, Theil–Sen consistently outperformed both bi-split and tri-split hand-fit lines in power and efficiency. Although many trend estimations are available, we use two best known and most accessible methods as exemplars for how to calculate trend line as the first step in GROT.

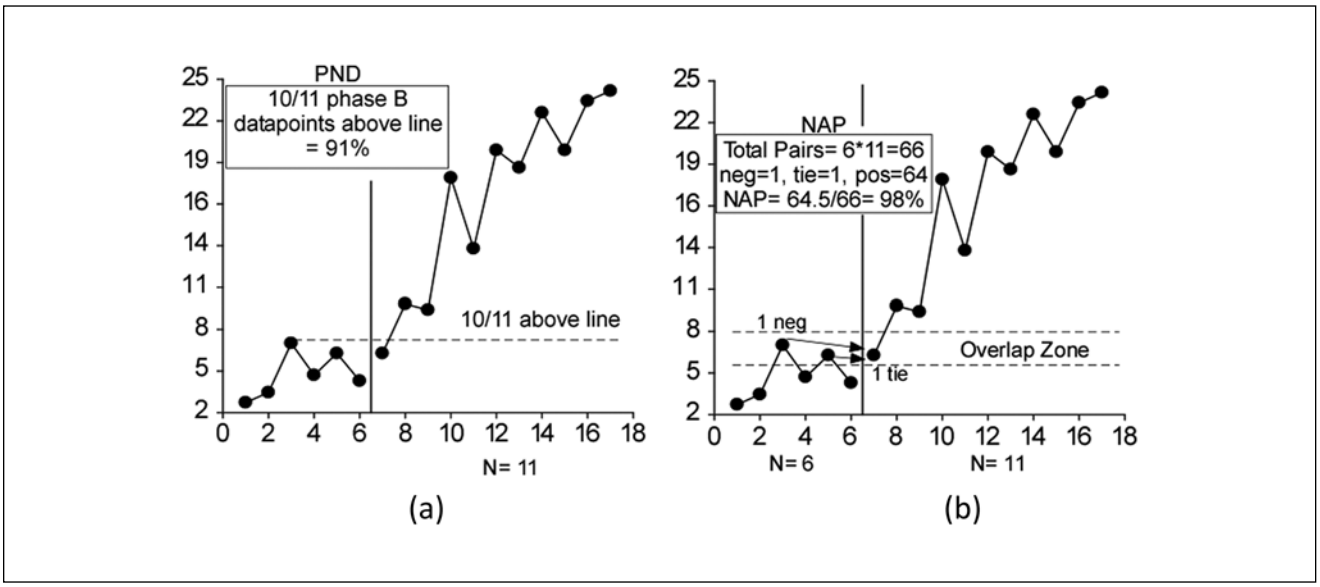

Figure 1. Example data set illustrating the calculations of (a) PND and (b) NAP.
Note. PND = percent of nonoverlapping data; NAP = nonoverlap of all pairs.

Two Nonoverlap Indices

Just as two "commonly used" and likely known trend methods were selected as examples, we provide illustrations for two overlap methods which are most commonly known PND (Scruggs et al., 1987) and NAP (Parker & Vannest, 2009). Just as GROT is a method that can include any type of method for calculating Phase A trend, GROT is applicable for any type of AB nonoverlap index. Simple AB contrasts are chosen for the current demonstration due to the fact that this contrast is the most basic SCR design to analyze. Part of the appeal of SCR for intervention researchers is the flexibility in constructing the design. Recent guidelines for evaluating SCR design quality (Kratochwill et al., 2010) call for a minimum of three demonstrations of experimental control. For example, this may take the form of an ABAB "reversal" design or three staggered AB contrasts within a multiple baseline design. It should be noted that the AB contrast alone does not meet minimum criteria for SCR design; however, multiple AB contrasts can be aggregated to provide an omnibus effect size for more complicated SCR designs.

The generality of GROT is demonstrated here by applying it with PND and NAP. These two nonoverlap indices are demonstrated by applying them first outside of GROT, to raw scores from the first example data set (see Figure 1).

In Figure 1, PND is calculated as the percentage of Phase B data points above the highest data point in Phase A. First, the number of data points in Phase B is noted (n_B = 11). Next, the highest point in Phase A is located (third data point is 6.8), and a horizontal line drawn to their right. Above that line, we count 10 of the 11 Phase B data points. PND is calculated as 10 / 11 = 91%. In Table 1, the third column shows the two scores critical to PND calculation: (a) the highest score in Phase A (6.8) and (b) the only smaller value in Phase B (6.1).

NAP is a "complete" nonoverlap method, as it equally considers all data points in both phases. As a complete method, it is supported by "dominance" statistics, mentioned earlier and described in more detail later in this article. NAP computation is not as simple as PND, so some users will prefer software, but hand calculation is easy enough to be accessible. NAP is output directly from a receiver operator characteristic (ROC) curve module as empirical AUC and may also be obtained from the two *U* values from a Mann–Whitney *U* test.

Although NAP can be instantly calculated by AUC or Mann–Whitney *U* test, hand calculation is described first to enhance understanding. The NAP formula is the number positives added to .5 the number of ties, minus the negatives, divided by the number of pairs: (positives + .5 × ties) / pairs.

First, the number of data points in Phases A and B are multiplied together to obtain the total number of paired comparisons (6 × 11 = 66 pairs). Next, the "overlap zone" is visually identified (see Figure 1b). The overlap zone extends from just under the lowest Phase B data point up to just above the highest Phase A data point, this zone will contain data to be labeled "negative" or a "tie." For simplicity, we count the negatives and subtract from total number of pairs to get the number of positives rather than count all positives, which is generally faster. Ties are data equal to each other on the *Y*-axis in Phases A and B. Figure 1b overlap zone contains

Table 1. Control of Phase A Tri-Split and Bi-Split Trends via Semipartialling in First Sample Data Set.

Time	Phase	Score	Tri-split slope × Time (.53)	Tri-split detrended	Bi-split slope × Time (.40)	Bi-split detrended
1	A	2.7	0.5	2.2	0.4	2.3
2	A	3.4	1.1	2.4	0.8	2.6
3	A	6.8	1.6	**5.2* PND**	1.2	**5.6* PND**
4	A	4.6	2.1	**2.5***	1.6	3.0
5	A	6.1	2.6	**3.5***	2.0	**4.1***
6	A	4.2	3.2	1.1	2.4	1.8
7	B	6.1	3.7	**2.4* PND**	2.8	**3.3* PND**
8	B	9.5	4.2	5.3	3.2	6.3
9	B	9.1	4.7	**4.4* PND**	3.6	**5.5* PND**
10	B	18.2	5.3	12.9	4	14.2
11	B	13.3	5.8	7.5	4.4	8.9
12	B	20.1	6.3	13.8	4.8	15.3
13	B	18.9	6.8	12.1	5.2	13.7
14	B	22.7	7.4	15.4	5.6	17.1
15	B	20.1	7.9	12.2	6	14.1
16	B	23.5	8.4	15.1	6.4	17.1
17	B	24.2	8.9	15.3	6.8	17.4
				NAP: pairs = 66, negative = 4, positive = 62, ties = 0; 62 / 66 = .94		NAP: pairs = 66, negative = 3, positive = 63, ties = 0; 63 / 66 = .95

Note. PND = percent of nonoverlapping data; NAP = nonoverlap of all pairs. Values in NAP's "overlap zone" are presented in bold and with asterisks.

one negative pair (Data Point 3 compared with Data Point 7; negative = 1) and one tie (Data Point 5 compared with Data Point 7; tie = .5). Note that this example has only one data point in Phase B for illustration, and any additional data points in the overlap zone would require additional comparisons. The same "overlap zone" is represented in Table 1, column 5, by data presented in bold and with asterisks. Out of 66 paired comparisons, negative = 1 and tie = .5, so the remaining pairs must be positive (positive = 64.5). Therefore, the PND [NAP = (positive + .5 × ties) / pairs] equals (64.5 + .5) / 66 = 98%. NAP calculations are more involved than for PND, so some users will prefer to obtain NAP directly as "empirical AUC" from a ROC test, also termed a *Diagnostic Precision* test or *Sensitivity/Specificity* test (Swets, 1995). A complete description of NAP is available in Parker and Vannest (2009) for interested readers.

Calculation with a stats package is straightforward. Input "Phase" as the actual or true value, and "Scores" as the criterion or test variable. The empirical AUC is output (.98), along with its statistical significance ($p < .00$), and 90% confidence intervals (CI) [.84, .99]. NAP also is available from a Mann–Whitney *U* test, with one simple calculation required. A full-featured Mann–Whitney module will output larger and smaller *U* values, which for these data are large or $U_L = 64.5$, and small or $U_S = 1.5$, so NAP = $U_L / (U_L + U_S) = 64.5 / 66 = .98$. The significance test yields $Z = 3.12$, so two-tailed $p = .001$. ROC-AUC and Mann–Whitney *U* tests yield identical NAP values, but they rely on different sampling distributions, so their *p* values and CIs will differ. For NAP, chance-level results are .50. NAP can be transformed so that chance-level results equal 0: $NAP_{0–100} = 1 - (NAP_{50–100} / .5)$. This transformation would change the NAP of .98 to .95.

The two nonoverlap results from original data (PND = 91% and NAP = 97.7%) will be compared with PND and NAP values obtained from the GROT baseline trend control procedure. Because PND is the simplest nonoverlap to calculate, and NAP is the most powerful, these two options will be used to demonstrate GROT.

GROT on First Example Data

GROT's four steps are demonstrated using the same example data set: (a) set a trend line to Phase A (this example will use a tri-split line); (b) keeping parallel to the original slope, move the trend line down to the intersect of *X*- and *Y*-axes, and also redraw the phase separation line perpendicular to the slope; (c) rotate the graph so the trend line now becomes a new horizontal axis; and (d) calculate nonoverlap. These steps are shown with the tri-split slope in Figure 2 and in Table 1. Two nonoverlap summaries, PND and NAP, are both calculated for sake of comparison.

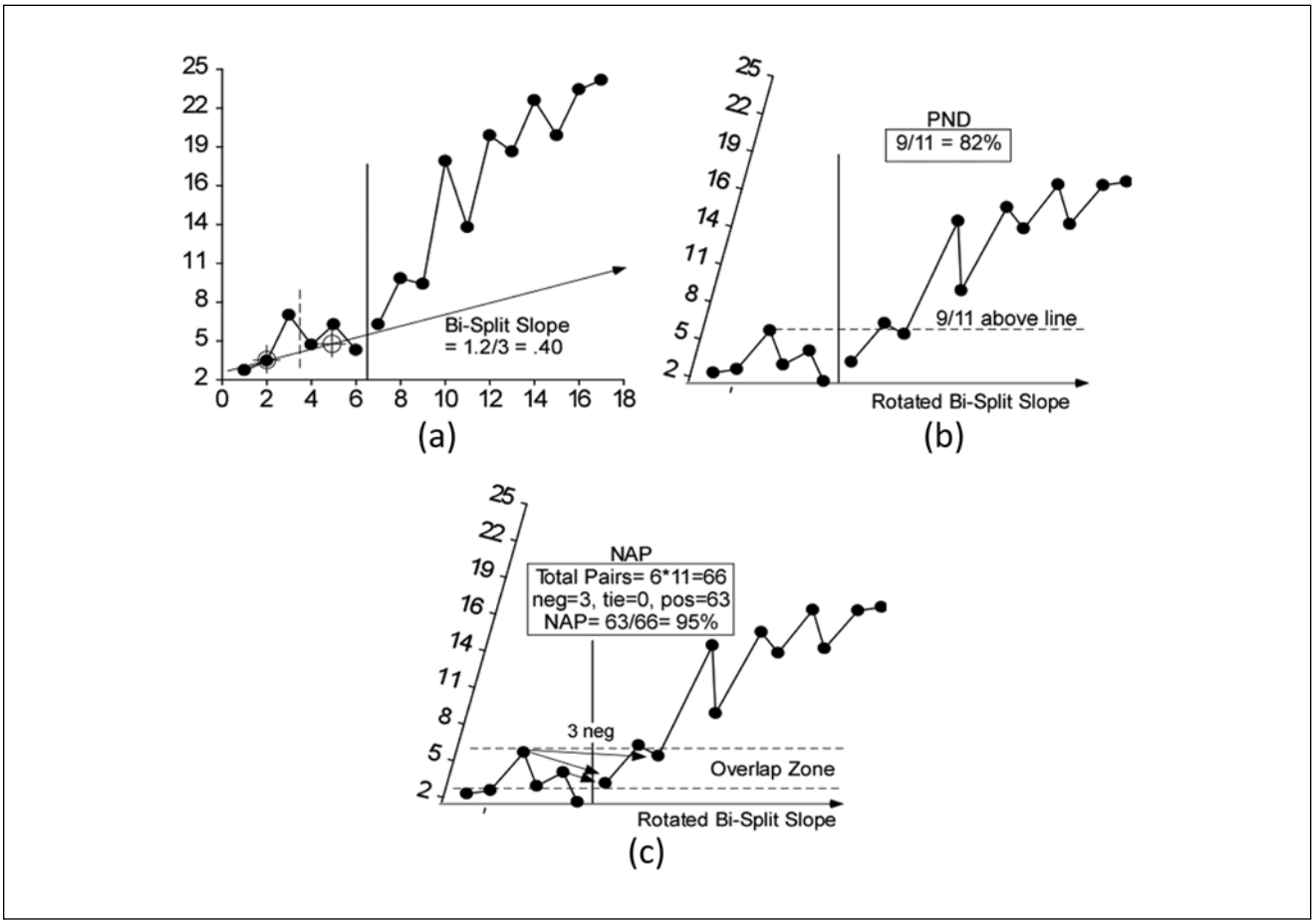

Figure 2. Example data set demonstrating (a) Koenig's bi-split line plotted for Phase A and extended through Phase B; (b) the bi-split line dropped to the *X*- and *Y*-axis intersect, the graph rotated, and PND recalculated; (c) recalculation of NAP on adjusted data.
Note. PND = percent of nonoverlapping data; NAP = nonoverlap of all pairs.

Figure 2 shows the Koenig bi-split slope (.40) plotted through the median intersects of the two halves of Phase A data. Figure 2b shows dropping the bi-split trend line down to run through the axis intersect, and then redrawing the phase division line perpendicular to the reset bi-split slope. Figure 2b also shows the final two steps: rotating the graph to use the bi-split trend line as the horizontal axis, and re-calculating PND. No redrawing of the graph is needed; it can be simply rotated (as was done here). PND calculated on the "detrended" data is 82%, which is less than the 91% calculated on the original data (see Figure 1), demonstrating that the nonoverlap effect was reduced by controlling the Phase A positive trend. This example shows that the original trended data and nonoverlap calculation overestimated the treatment effects. This visual-graphic method works equally well with any nonoverlap statistic. For example, Figure 2c shows NAP recalculated on the GROT detrended data. In this example, the original effect size estimate of 98% is reduced to 95%.

GROT validation by Allison and Gorman regression control on first data set. GROT accuracy (as a hand-calculated technique) can be validated by comparing its results with those from the best available regression control method by Allison and colleagues (Faith et al., 1996). Their procedure is as follows: (a) Calculate Phase A slope (they use a regression slope, but we will use the tri-split slope = .53); (b) multiply the slope by a simple linear series (see Table 1, column 4); and (c) subtract the series of those products from the original data series. This will result in transformed scores, with Phase A trend removed (see Table 1, column 5). In column 5, key scores for calculating PND are labeled PND, and key scores for calculating NAP are asterisked. For PND, the highest Phase A score (5.2) is identified, and all but two Phase B scores (2.4, 4.4) are higher, so PND = 9 / 11 = 82%. For NAP, the "overlap zone" contains three Phase A scores (5.2, 2.50, 3.5) and two Phase B scores (2.4, 4.4), and of their pairings, four are in the negative direction (negative = 4), and ties = 0, so of the 66 total pairs, 62 must be positive

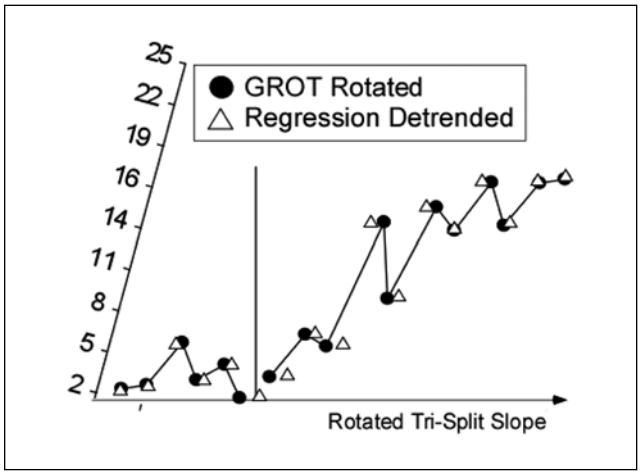

Figure 3. Comparison of GROT-rotated scores with regression (Allison & Gorman, 1993) detrended scores.
Note. GROT = graph rotation for overlap and trend.

(positive = 62). Therefore, NAP = 62 / 66 = 94%. These results are identical to those obtained from the GROT-rotated graphs.

A graphic comparison also validates GROT against the Allison and Gorman control method. In Figure 3, the rotated and the Allison and Gorman detrended (semipartialled) scores are plotted together.

Their respective locations on the *Y*-axis are identical. Their respective locations on the *X*-axis are not identical, because rotating the graph skews the *X*-axis. However, a nonoverlap test is ordinal, so it does not depend on correct intervals on the *X*-axis—only correct order, or relative positions. Thus, this visual-graphic test also validates GROT.

GROT was also used with a tri-split median trend from Phase A, demonstrated in Table 1, columns 6 and 7. In column 6, the bi-split trend (.40) is multiplied by the time series, and the result is subtracted from original scores to yield detrended scores in column 7. Again, critical values for calculating PND are labeled, and values in NAP's "overlap zone" are asterisked. PND again equals 9 / 11 = 82%. NAP is slightly higher (than for the tri-split slope): 63 / 66 = 95%. These results are identical to those obtained from visual analysis of the GROT-rotated graph in Figures 2b and 2c.

GROT on a second example data set. GROT is applied to control positive baseline trend in a second demonstration data set. Figures 4a and 4b show PND and NAP calculation on the original data. Figure 4c shows the Tukey tri-split slope calculated for Phase A and extended through Phase B. Figures 4d and 4e show the recalculation of PND and NAP on the GROT-rotated data. PND is reduced from 90% to 80% due to the GROT rotation, and the more comprehensive analysis NAP is reduced from 99% to 92%.

GROT validation by Allison and Gorman on the second example data set. Figure 4f repeats for the second example data set, the validation of GROT by the Allison and Gorman semipartialling regression procedure. Original and detrended scores for the second example are presented in Table 2 to validate the graph rotation procedure. Considering first the tri-split slope (columns 4 and 5), for PND, the highest detrended score is the eighth in order (7.2). In Phase B (column 5), 8 of 10 scores are higher than 7.2, so PND = 80%. For NAP, the scores in the "overlap zone" are all bold. Of these 6 × 2 = 12 combinations, 7 are "negatives" or drop from Phase A to Phase B. There are a total of 9 × 10 = 90 pairwise combinations between phases. With 7 paired comparisons negative, and no ties, the remainder must be positive. Therefore, NAP = (positive + .5 × ties) / pairs, which is 83 / 90 = 92%. The sixth and seventh columns in Table 2 show calculation of PND and NAP on GROT-rotated data using a bi-split trend line. There is no Figure associated with these columns; they are included to permit replication.

GROT Validation by Visual Analysis

GROT is designed to be a technique compatible with visual analysis, yet it was unknown whether the rotation would challenge the visual analyst, as this novel rotated graph may not have been previously encountered. Therefore, two questions were posed related to the interpretability of rotated graphs. The first question was whether visual analysts could make reliable judgments about behavior change from rotated graphs. The second question was how well visual analysts could identify the decreased behavior change from Phase A to B. The point of GROT is to display two phases with smaller differences in performance due to Phase A trend control. It was hypothesized that visual analysts would be able to detect those differences, at least in data sets with pronounced initial Phase A trends. Our hypotheses were that (a) rater agreement would be at least as high with GROT graphs as with original data graphs because of the effect of having a transformed flat baseline and (b) that raters would correctly detect less change from Phase A to B in some GROT graphs and not in others, but would not identify *more* change in GROT graphs.

Method

From a corpus of 372 published single-case distinct data series, AB data series were identified which met the dual criteria of (a) visually apparent Phase A trend in the same direction as desired behavior change and (b) this positive Phase A trend confirmed by Kendall's Tau rank correlation test. A total of 49 AB data sets met these criteria. For each of the 49 data sets, two graphs were presented on separate 4 × 8 note cards, first with the original data graph, second with a GROT-rotated graph. Three graduate students in

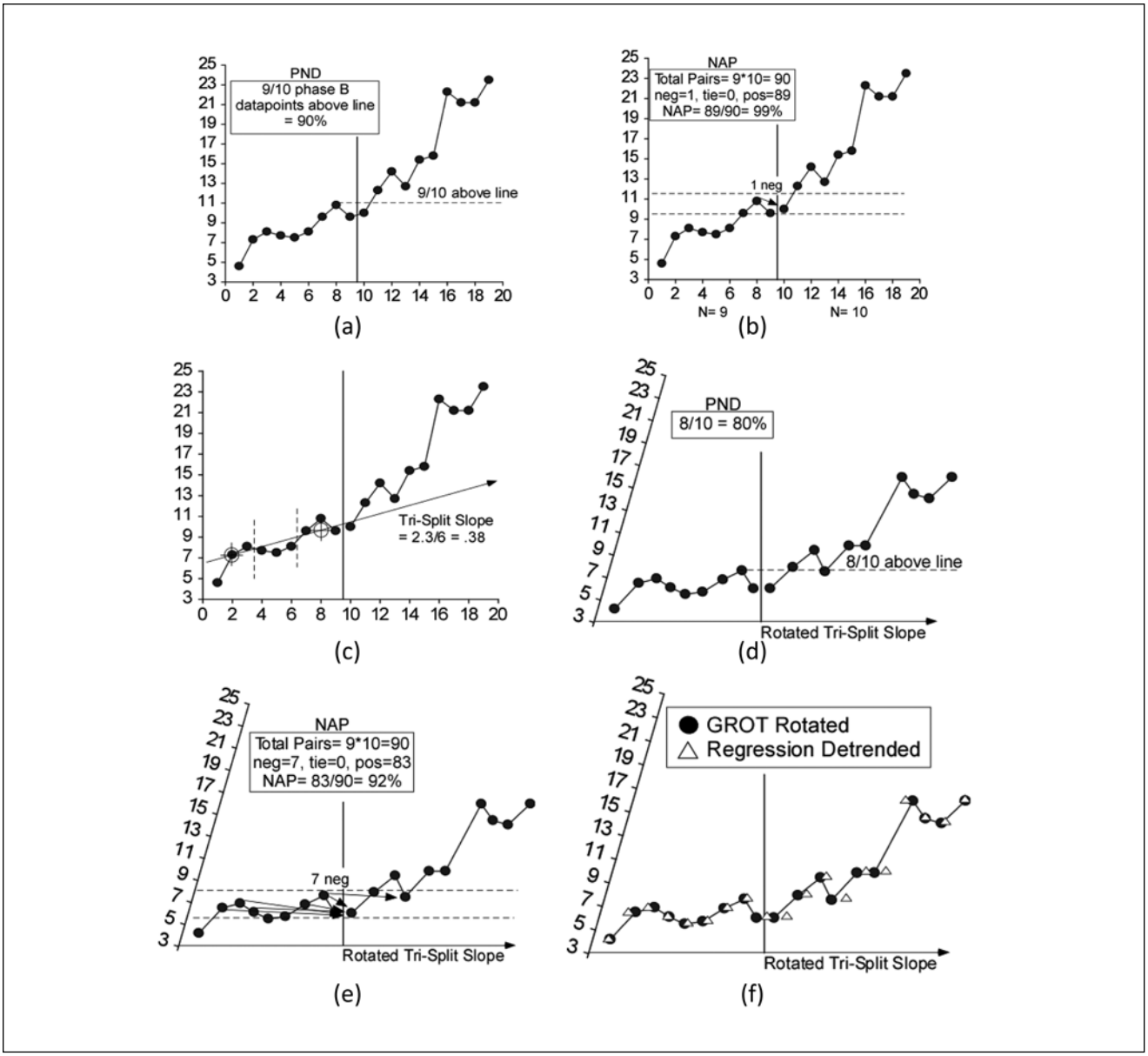

Figure 4. Second example data set (a) uncorrected PND calculation; (b) uncorrected NAP calculation; (c) Phase A trend plotted as tri-split slope; (d) GROT rotated, and PND calculated on rotated scores; (e) GROT rotated, and NAP calculated on rotated scores; and (f) validation of GROT-rotated data by Allison et al. regression detrending.
Note. PND = percent of nonoverlapping data; NAP = nonoverlap of all pairs; GROT = graph rotation for overlap and trend.

school psychology and special education independently rated each graph for magnitude of change from Phase A to B on a 3-point scale: *large, medium*, and *small*. Graphs were presented in random order.

To answer the first question about reliable judgments, interrater reliabilities were calculated and compared among the three raters on the original data graphs and the GROT graphs. To answer the second question about visual judges' ability to judge GROT graphs as showing smaller change, we calculated cross-tabulations for each rater on original graph ratings versus GROT graph ratings. The resulting matrices were then examined for shifts in amount and direction of perceived effects from original to GROT graphs. Note that the two graph types were presented randomly, not in pairs, to reduce bias.

Results

To answer the first question about reliable judgments, linear weighted Cohen's kappa (κ-LW) was calculated by Richard

Table 2. Control of Phase A Koenig Bi-Split Trend via Semipartialling in Second Sample Data Set.

Time	Phase	Score	Tri-split slope × Time (.53)	Tri-split detrended	Bi-split slope × Time (.40)	Bi-split detrended
1	A	4.6	0.4	4.2	0.4	4.2
2	A	7.3	0.9	**6.4***	0.8	**6.5***
3	A	8.1	1.3	**6.8***	1.3	**6.8***
4	A	7.7	1.8	**5.9***	1.7	**6.0***
5	A	7.5	2.3	5.3	2.1	5.4
6	A	8.1	2.7	5.4	2.5	5.6
7	A	9.6	3.2	**6.5***	2.9	**6.7***
8	A	10.8	3.6	**7.2* PND**	3.4	**7.4* PND**
9	A	9.6	4.1	**5.6***	3.8	**5.8***
10	B	10	4.5	**5.5* PND**	4.2	**5.8* PND**
11	B	12.3	4.9	7.4	4.6	7.7
12	B	14.2	5.4	8.8	5.0	9.2
13	B	12.7	5.9	**6.9* PND**	5.5	**7.2* PND**
14	B	15.4	6.3	9.1	5.9	9.5
15	B	15.8	6.8	9.0	6.3	9.5
16	B	22.3	7.2	15.1	6.7	15.6
17	B	21.2	7.7	13.5	7.1	14.1
18	B	21.2	8.1	13.1	7.6	13.6
19	B	23.5	8.6	15.0	8.0	15.5
				NAP: pairs = 90, negative = 7, ties = 0; 83 / 90 = .92		NAP: pairs = 90, negative = 7, ties = 0; 83 / 90 = .92

Note. PND = percent of nonoverlapping data; NAP = nonoverlap of all pairs. Values in NAP's "overlap zone" are presented in bold and with asterisks.

Table 3. Response Shifts From Three Judges Rating AB Graphs Before and After GROT Transformations.

	Raw data graphs								
	Rater A			Rater B			Rater C		
GROT graphs	Small	Medium	Large	Small	Medium	Large	Small	Medium	Large
Small	22	13	3	22	10	2	21	8	2
Medium	—	9	5	—	11	6	—	14	5
Large	—	—	7	—	—	8	—	—	9

Note. GROT = graph rotation for overlap and trend.

Lowry's (2011) open source kappa calculation webpage (http://faculty.vassar.edu/lowry/kappa.html). Kappa-LW, unlike simple kappa, is sensitive to the amount or degree of disagreement on an ordinal scale (Parker, Vannest, & Davis, in press). For the original graph judgments, kappa-LW among the three raters was .84, .70, and .63, and the corresponding simple percent agreements were .88, .76, and .71. Interrater agreement on the GROT graphs by kappa-LW was .82, .78, and .76, with simple percent agreements .88, .86, and .85. Therefore, the GROT graphs permitted slightly higher agreement among raters, likely due to their flat baselines.

The second question related to visual analysts' ability to detect smaller effects in the GROT graphs was answered by three cross-tabulations. Each rater (see Table 3) shows response shifts. Rater A judged 38 graphs as unchanged and 21 GROT graphs as showing reduced effects. Rater B judged 41 graphs as unchanged and 18 GROT graphs with smaller effects. Rater C judged 44 graphs as unchanged and 15 GROT graphs with smaller effects. There were no response shifts from original to GROT graphs indicating greater perceived effects.

Discussion

Nine overlap techniques are currently available to calculate standardized scores for determining the "size" of change between two or more phases (Parker et al., 2011). However, only two are capable of handling trend Tau-U and ECL. This is unfortunate and important because trend in baseline

data weakens conclusion validity (Kane, 2001; Kazdin, 2003; Orme, 1991) and thus our ability to promulgate practices with empirical evidence. Although many statistical models address this issue, they are largely inappropriate or inaccessible for nonparametric data or visual analysts. ECL, which was known as a technique in other fields in the 1940s, has a long history of use but limitations. Three in particular, the interpretation as a ratio of data points, low power, and inability to be applied universally are challenges that if overcome, would move the field forward and extend our knowledge base of valid, reliable techniques to address trend in nonoverlap analysis, which are compatible with visual analysis and avoid sophisticated statistical packages.

This article presented the GROT method for controlling positive baseline trend within a nonoverlapping data analysis and demonstrated validity by performance equal to the best, current regression method by Allison and colleagues (Allison & Gorman, 1993; Faith et al., 1996). This equivalence was demonstrated both numerically and graphically. GROT also demonstrated reliability with visual analyst ratings despite producing a novel rotated graph, which may have presented challenges. Finally, rater response shifts from original to GROT graphs indicate compatibility with visual judgments and produced reliable detection of the amount of change in simple AB graphs, more so than for original graphs.

GROT advances our knowledge base by providing an additional technique for visual analysts, a technique which appears to improve accuracy in determining effects. Over the past few decades, there has been ongoing research on the training and supports that would enhance reliability of visual judgments from graphs (Ferron & Jones, 2006; Fisher, Kelley, & Lomas, 2003; Ximenes, Manolov, Solanas, & Quera, 2009). The GROT graph may be of service toward that goal. GROT may have use as a visual analyst tool alone, aside from nonoverlap calculations.

Ratings of original and GROT graphs consistently placed GROT effects equal or lower than for the original graphs. No GROT graphs were identified as showing larger effects, and this strong finding held over three independent raters and 59 graphs. However, about 70% of graphs were judged as showing no change in magnitude of effect. There are several possible explanations. In some instances, we suspect that visual judges cannot detect small changes. Our 3-point scale of "smaller, same, larger" was less sensitive than necessary to detect changes. Or, in some instances, adjusting positive baseline trend may not have eliminated effects of a very large magnitude. However, the minimum amount of change detectable from original to GROT graphs is a question with practical implications for visual analysts as is an empirical comparison between effect size changes and visual analysis estimations.

Another way GROT advances in the field is its convenience. GROT is a method which can be carried out entirely with pencil and ruler on a paper graph, so it is fully accessible to visual analysis. It advances the field because nonoverlap is widely used by SCR practitioners but is commonly criticized for failing to consider positive "preexisting" Phase A trend. The addition of Phase A trend control permits nonoverlap methods to compete with leading parametric methods.

An asset of GROT is that it is a general graphic approach which works equally well for any trend line, including linear regression, Tukey tri-split, Koenig bi-split, or Theil–Sen slope. The Tukey tri-split and Koenig bi-split alternatives were demonstrated. GROT is applicable with any nonoverlap method. Here, only PND and NAP were applied, but other nonoverlap indices could be used as well.

Cautions on Controlling Baseline Trend

Although this article has the primary goal of promoting a new analytic method, it also needs to raise concerns about the overuse of baseline trend control, that is, its use with unreliable, highly variable trend lines. All trend control methods noted in this article, including ECL, Allison and Gorman's (1993) regression method, and GROT, adjust the full data series according to the slope of the Phase A trend line. However, there are at least three concerns with the use or misuse of such control. These concerns are not new (Scruggs & Mastropieri, 1994, 1998), but to date have not been addressed, so practitioners should be aware.

When Phase A trend is stable (i.e., lacks variability), a linear trend line is a reliable summary of the data points contained within the phase. Controlling the Phase A trend in cases with stable (i.e., nonvariable) data is likely to render a more appropriate estimate of effect across phases. In contrast, the application of baseline trend control with highly variable Phase A data raises some concerns. As the control is based on slope or trend line, it is blind to the potential unreliability of the Phase A trend line, so control from highly unreliable trend will have the same impact on Phase B scores as from a reliable trend. This is counterintuitive; some linear trend lines reflect data so poorly that they should not be fit to the data, let alone permitted to modify Phase B scores. Phase A data may simply lack linearity, and a straight line would be inappropriate.

The second concern is that control of Phase A trend becomes more extreme with a longer Phase B. Phase A trend line slope is most reliable within Phase A, and even more so at the center of Phase A. If extended through a long Phase B, the reliability or credibility of this slope quickly approaches zero. Given a long enough Phase B, the baseline trend control will transform Phase B data far outside the bounds of the score scale. Regression texts warn us about

the very low reliability of projections into the future, and the problem is even greater for $N = 1$ single-participant data.

The third concern is the open question of whether Phase A trend would continue unabated through Phase B had there been no intervention. This is a difficult question to answer statistically, but we do know that a strong trend in the first 5 or 8 data points of a baseline is not a good predictor of trend in the next 5 or 8 data points (Parker, Cryer, & Byrns, 2006). The evidence indicates that strong trend in the first third or half of a baseline tends to moderate considerably in the final two thirds or half of that same baseline. This evidence is not conclusive, but is suggestive that positive baseline trend may not continue at strength into Phase B.

We believe that these three concerns have sufficient weight that control of baseline trend should be exercised cautiously, which is to say not with quite short baselines or with highly variable baseline data. Baseline control has logical appeal and is carried out with precision. However, data transformations from short and highly variable Phase A data may be an exercise in false precision.

In summary, this article has presented and a visual-graphic method of controlling for baseline trend within data nonoverlap and included preliminary reliability and validity comparisons. The method offers greater flexibility and power than White and Haring's (1980) respected ECL method and is more accessible and more directly interpretable than the Allison and Gorman's (1993) regression method. GROT leads to reliable judgments of behavior change and correctly reflects reduced effects from original data. Initial indications of performance are positive, however, as with any new analytic method, it requires testing over time and by a variety of researchers. If present results are borne out by others, then we hope to have contributed to the merging of statistical and visual analysis of SCR data.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Abramson, J. H. (2010). Programs for epidemiologists–Windows version (WinPepi) [Computer software]. Retrieved from http://www.brixtonhealth.com/pepi4windows.html

Acion, L., Peterson, 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. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behavior, Research, and Therapy, 31*, 621–631.

Bartlett, M. S. (1949). Fitting a straight line when both variables are subject to error. *Biometrics, 5*, 207–212.

Brown, G. W., & Mood, A. M. (1951). On median tests for linear hypotheses. *In Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press.

Calkin, A. B. (2005). Precision teaching: The standard celeration charts. *The Behavior Analyst Today, 6*, 207–213.

Center, B. A., Skiba, R. J., & Casey, A. (1985–1986). A methodology for the quantitative synthesis of intrasubject design research. *Journal of Special Education, 19*, 387–400.

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

Crosbie, J. (1993). Interrupted time-series analysis with brief single-subject data. *Journal of Consulting and Clinical Psychology, 61*, 966–974.

D'Agostino, R. B., Campbell, M., & Greenhouse, J. (2006). Non-inferiority trials: Continued advancements in concepts and methodology. *Statistics in Medicine, 25*, 1097–1099.

Delaney, H. D., & Vargha, A. (2002). Comparing several robust tests of stochastic equality with ordinally scaled variables and small to moderate sized samples. *Psychological Methods, 7*, 485–503.

Faith, M. S., Allison, D. B., & Gorman, B. S. (1996). Meta-analysis of single-case research. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 245–277). Mahwah, NJ: Lawrence Erlbaum.

Ferron, J., & Jones, P. K. (2006). Tests for the visual analysis of response-guided multiple-baseline data. *Journal of Experimental Education, 75*, 66–81.

Fisher, W. W., Kelley, M. E., & Lomas, J. E. (2003). Visual aids and structured criteria for improving visual inspection and interpretation of single-case designs. *Journal of Applied Behavior Analysis, 36*, 387–406.

Granato, G. E. (2006). Kendall–Theil Robust Line (KTRLine—version 1.0) [Computer software]. Retrieved from http://pubs.usgs.gov/tm/2006/tm4a7/

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

Hintze, J., & Stecker, P. (2006). *Data-based instructional decision making* [Online PowerPoint presentation]. Retrieved from http://www.studentprogress.org/summer_institute/rti/DataBasedInstructionalDecisionMaking/DataBasedInstructionalDecisionMaking_powerpoint.ppt

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). Understanding robust and exploratory data analysis. New York, NY: John Wiley.

Hodges, J. L., Jr., & Lehmann, E. L. (1956). The efficiency of some nonparametric competitors of the t-test. *The Annals of Mathematical Statistics, 27*, 324–335.

Hollander, M., & Wolfe, D. A. (1999). *Nonparametric statistical methods* (2nd ed.), New York, NY: John Wiley.

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

Johnstone, I., & Velleman, P. F. (1985). Efficient scores, variance decompositions, and Monte Carlo swindles. *Journal of the American Statistical Association, 80*, 851–862.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319–342.

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

Kazdin, A. E. (2003). *Research design in clinical psychology* (4th ed.). Needham Heights, MA: Allyn & Bacon.

Koenig, C. (1972). *Charting the future course of behavior*. Kansas City, KS: Precision Media.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case designs technical documentation*. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Lindsley, O. R. (1991). Precision teaching's unique legacy from B. F. Skinner. *Journal of Behavioral Education, 1*, 253–266.

Lowry, R. (2011). Kappa as a measure of concordance in categorical sorting. [Web-based software]. Retrieved from http://faculty.vassar.edu/lowry/kappa.html

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.

Nair, K. R., & Srivastava, M. P. (1942). On a simple method of curve fitting. *Sankhayā: Indian Journal of Statistics, 6*, 121–132.

O'Brien, S., & Repp, A. C. (1990). Reinforcement-based reductive procedures: A review of 20 years of their use with persons with severe or profound retardation. *Journal of the Association for Persons With Severe Handicaps, 15*, 148–159.

Orme, J. G. (1991). Statistical conclusion validity for single-system designs. *Social Service Review, 65*, 468–491.

Parker, R., Stein, M., & Tindal, G. (1992). Estimating trend in progress monitoring data: A comparison of simple line-fitting methods. *School Psychology Review, 2*, 300–313.

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. J. (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. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy, 40*, 357–367.

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

Parker, R. I., Vannest, K. J., & Davis, J. L. (2011). Effect size in single case research: A review of nine non-overlap techniques. *Behavior Modification, 35*(4), 303-322. doi:10.1177/0145445511399147

Parker, R. I., Vannest, K. J., & Davis, J. L. (in press). Reliability for multi-category rating scales. *Journal of School Psychology, 35*, 302–322.

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining non-overlap and trend for single case research: Tau-U. *Behavior Therapy, 42*, 284–299.

Parsonson, B. S., & Baer, D. M. (1978). The analysis and presentation of graphic data. In T. Kratochwill (Ed.), *Single subject research*. New York, NY: Academic Press.

Pennypacker, H. S., Koenig, C. H., & Lindsley, O. R. (1972). *Handbook of the standard behavior chart*. Kansas City, KS: Precision Media.

Scotti, J. R., Evans, I. M., Meyer, L. H., & Walker, P. (1991). A meta-analysis of intervention research with problem behavior: Treatment validity and standards of practice. *American Journal on Mental Retardation, 96*, 233–256.

Scruggs, T. E., & Mastropieri, M. A. (1994). The utility of the PND statistic: A reply to Allison and Gorman. *Behavior, 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., & Casto, G. (1987). The quantitative synthesis of single subject research: Methodology and validation. *Remedial and Special Education, 8*, 24–33.

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. *American Statistical Association Journal, 63*, 1379–1389.

Sokal, R. R., & Rohlf, F. J. (1995). *Biometry*. New York, NY: W.H. Freeman.

Swets, J. A. (1995). *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. Mahwah, NJ: Lawrence Erlbaum.

SYSTAT (2008). Mystat Software (Version 11) [Computer software]. Retrieved from http://www.systat.com/

Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis, III. *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen A, 53*, 1397–1412.

Tukey, J. W. (1977). *Exploratory data analysis*. Menlo Park, CA: Addison-Wesley.

Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *Annals of Mathematical Statistics, 11*, 284–300.

White, O. R. (1974). *The split middle: A quickie method of trend estimation* (3rd rev.). Unpublished manuscript, University of Washington, Experimental Education Unit, Child Development and Mental Retardation Center, Seattle, WA.

White, O. R. (1986). Precision teaching—Precision learning. *Exceptional Children, 52*, 522–534.

White, O. R., & Haring, N. G. (1980). *Exceptional teaching* (2nd ed.). Columbus, OH: Merrill.

White, O. R., Rusch, F. R., Kazdin, A. E., & Hartmann, D. P. (1989). Applications of meta-analysis in individual subject research. *Behavioral Assessment, 11*, 281–296.

Wilcox, R. R. (2010). *Fundamentals of modern statistical methods: Substantially improving power and accuracy* (2nd ed.). New York, NY: Springer.

Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *Journal of Special Education, 44*, 18–28.

Ximenes, V. M., Manolov, R., Solanas, A., & Quera, V. (2009). Factors affecting visual inference in single-case designs. *Spanish Journal of Psychology, 12*, 823–832.