Available online at www.sciencedirect.com
ScienceDirect
Behavior Therapy xx (2011) xxx–xxx
Behavior Therapy
www.elsevier.com/locate/bt

Combining Nonoverlap and Trend for Single-Case Research: Tau-U

Richard I. Parker
Kimberly J. Vannest
John L. Davis
Stephanie B. Sauber
Texas A&M University at College Station

A new index for analysis of single-case research data was proposed, Tau-U, which combines nonoverlap between phases with trend from within the intervention phase. In addition, it provides the option of controlling undesirable Phase A trend. The derivation of Tau-U from Kendall's Rank Correlation and the Mann-Whitney U test between groups is demonstrated. The equivalence of trend and nonoverlap is also shown, with supportive citations from field leaders. Tau-U calculations are demonstrated for simple AB and ABA designs. Tau-U is then field tested on a sample of 382 published data series. Controlling undesirable Phase A trend caused only a modest change from nonoverlap. The inclusion of Phase B trend yielded more modest results than simple nonoverlap. The Tau-U score distribution did not show the artificial ceiling shown by all other nonoverlap techniques. It performed reasonably well with autocorrelated data. Tau-U shows promise for single-case applications, but further study is desirable.

NONOVERLAP MODELS VERSUS REGRESSION MODELS

Single-case research (SCR) has received renewed interest in the behavioral sciences for its focus on change within an individual rather than change in the group aggregate (Borckardt et al., 2008). Statistical analysis for evaluating change in SCR designs are still in an early stage of development. Ordinary least squares regression analysis (OLS) with a long history of use in large *N* studies, has shown unequalled flexibility and power when applied to SCR data (Allison & Gorman, 1993; Busk & Serlin, 1992; Parker & Brossart, 2003). However, OLS has been criticized for failing to address the unique constraints of short time series data that are typical in SCR (Parsonson & Baer, 1992; Scruggs & Mastropieri, 1994). OLS is a parametric statistical test, and as such requires a normal score distribution, constant variance, and interval level measurement. Applying OLS to SCR data has been criticized because these data often do not meet OLS assumptions of constant variance, normality, and linearity of relationship, and the scaling assumption of at least an interval-level scale (Cohen & Cohen, 1983; Kutner, Nachtsheim & Neter, 2004). These problems notwithstanding, only OLS analysis has to date been able to demonstrate (a) control of undesirable positive baseline trend; (b) sensitivity to improvement in level change trends; (c) adequate power for short data series; and (d) the ability to discriminate well among published data sets, avoiding ceiling or floor effects. All nonoverlap indices suffer from a ceiling effect of 100%; they are insensitive to amount of separation of data contrasted between two phases beyond the point where there is no overlap.

At least four regression models have been designed to do those four things, which are summarized in texts by Franklin, Allison, and Gorman (1997), and Kratochwill and Levin (1992). They are (a) Crosbie's ITSACORR model (1993, 1995); (b) the Last Treatment Day prediction

Address correspondence to Richard I. Parker, Ph.D., Texas A&M University, 604 Harrington Office Building, Mail Stop 4225, College Station, TX 77843; e-mail: rparker@tamu.edu.

0005-7894/xx/xxx–xxx/$1.00/0

Please cite this article as: Richard I. Parker, et al., Combining Nonoverlap and Trend for Single-Case Research: Tau-U, *Behavior Therapy* (2011), 10.1016/j.beth.2010.08.006

technique of White, Rusch, Kazdin, and Hartmann (1989); (c) Center, Skiba, and Casey's (1985–1986) mean-shift and mean-plus-trend family of models; and (d) Allison et al.'s mean-shift and mean-plus-trend models (Allison & Gorman, 1993; Faith, Allison, & Gorman, 1997).

Crosbie's ITSACORR (1993, 1995) was positively cited by several researchers for a decade, but used infrequently, and has suffered two major setbacks. First, the experience of several researchers, including ourselves, was that its results bore little relationship to those from other models. Furthermore, ITSACORR results were not substantiated by visual analysis. Finally, the statistician Brad Huitema (2004) described "fatal flaws" in the model, in response to which Crosbie officially retired it: "I trust Brad's scholarship, so effective immediately ITSACORR is officially retired…. Now it's dead, and will soon be replaced" (Southerly, 2006).

The last treatment day (LTD) prediction technique of White et al. (1989) extended the baseline trend clear to the end of the treatment period; the "last treatment day." The predicted value at LTD was differenced from the LTD value predicted (as a Y_{hat} score) from the Phase B trend line alone, and the two predicted values at LTD were subtracted. The standard error of the difference was calculated for the two predicted scores. A Cohen's *d* effect size was then calculated from their difference divided by the pooled standard error term. Two flaws of this model were (a) linear prediction from Phase A to the end of Phase B resulted in extreme scores and extreme differences, and therefore, extreme effect sizes; and (b) the statistical power of the technique was quite weak due to the large error involved in predicting an individual score far into the future, to the end of Phase B (Parker & Brossart, 2006). Applied regression texts commonly warn that prediction of scores into the future is hazardous, even with large data sets and short-term predictions (Neter, Kutner, Nachtsheim, & Wasserman, 1996).

Center et al.'s (1985–1986) method marked a new level of sophistication, including both mean shift and trends in a single index, while controlling trend. However, in attempting to control positive baseline trend, Center's method also undesirably controlled some trend from the intervention phase. Center's method was critiqued and improved on by Allison, Faith, and colleagues (Allison & Gorman, 1993; Faith et al., 1997), whose model controlled only Phase A trend, but extended through the entire data series. The Allison et al. model is considered by most to be the leading OLS approach. It has proved itself in several published studies and at least two meta-analyses (Allison & Gorman, 1993).

Nonoverlap or "dominance" (Sprent & Smeeton, 2007) indices of improvement are based on comparisons of individual data points across two groups (two phases). Nonoverlap does not summarize the difference between central tendency (mean, median, or mode), but rather the separation of the two "data clouds," giving equal attention to all data points. The "dominance" of one data cloud over another is its degree of elevation above the other on a vertical score axis. Judging data overlap between phases has been a part of visual analysis since at least the 1960s, along with judging data trend (Cooper, Heron, & Heward, 1987; Johnston & Pennypacker, 1993; Kazdin, 1982). Nonoverlap was first measured statistically in the mid-1980s (Scruggs, Mastropieri, & Casto, 1987), and in the past two decades nonoverlap techniques have increased in number and refinement (Parker & Vannest, 2009; Parker, Vannest, & Brown, 2009). Nonoverlap methods vary mainly in how ties (across phases) are handled, and how overlapping versus nonoverlapping data pair counts are combined. However, all *complete* nonoverlap indices have in common the pairwise comparison of individual data points across Phases A and B, to determine "dominance" of one score set over the other (Cliff, 1993). The most recently published nonoverlap method, termed NAP (nonoverlap of all pairs) can be derived from Sommer's *d*, or from a receiver operator curve (ROC) analysis as area under the Curve (AUC; Parker & Vannest, 2009). NAP equals percent of nonoverlapping data. The new method demonstrated in this paper, derived from Kendall's Rank Correlation (Kendall & Gibbons, 1999), is the percent of nonoverlapping data minus the percent of overlapping data. In other respects, NAP and this new method are equivalent.

Besides its long use as part of visual analysis, and its user-friendliness (often carried out with pencil and ruler), nonoverlap has other strengths. First, nonoverlap methods are "distribution free," not requiring interval-level measurement or a linear relationship between time and scores, nor requiring constant variance or a normal distribution (Armitage, Berry, & Matthews, 2002). Nonoverlap methods also are robust or resistive to the undue influence of outlier scores, a particular strength in client-based research where "bouncy" scores are common. Furthermore, in some data sets the nonoverlap or "dominance" of one phase over another is a better, more sensitive summary than is mean or even median difference (Cliff, 1993; Huberty & Lowman, 2000). When scores are severely skewed, are bi- or tri-modal, or otherwise lack central tendency, a mean or median is not a good distribution summary (Wilcox, 2001).

Please cite this article as: Richard I. Parker, et al., Combining Nonoverlap and Trend for Single-Case Research: Tau-U, *Behavior Therapy* (2011), 10.1016/j.beth.2010.08.006

In those cases it makes more sense to consider all data points equally, as a dominance summary does.

Although nonoverlap methods are "distribution free" (Cliff, 1993), they are held to the standard of serial independence or lack of autocorrelation (r_{auto}), which applies to residual scores from an analysis. The serial independence requirement makes the exception for linear trend (which is 100% autocorrelated), as that is desired and expected in most time series data (Neter et al., 1996). The best evidence to date is that one third or more of published data sets from SCR designs are positively autocorrelated to an undesirable degree, with $r_{auto} > .20$ or .25, regardless of p value (Matyas & Greenwood, 1996; Parker, Cryer, & Byrns, 2006; Sharpley & Alavosius, 1988; Suen & Ary, 1989). Data r_{auto} is an important consideration in SCR data analysis and should be calculated and controlled. Several methods are currently available to accomplish this task, including simulating r_{auto} in a data set and testing its significance through resampling or bootstrap (Borckardt et al., 2008); however, the best established method for controlling r_{auto} is back casting with an autoregressive integrated moving average model (ARIMA) AR1 (1, 0, 0) model (Box & Jenkins, 1976; Glass, Willson, & Gottman, 1975; Jones, Vaught, & Weinrott, 1977). ARIMA finds solutions iteratively through maximum likelihood methods. But because ARIMA can be cumbersome, it is rapidly being replaced by methods seamlessly integrated into regression software. SAS software contains no less than 10 such methods, and four of the most popular methods are recently included in the "GNU Regression, Econometric and Time-Series Library" (GRETL; Cottrell & Lucchetti, 2009) software, freely downloadable from http://gretl.sourceforge.net/. Among the most favored is the generalized least squares Prais–Winsten (Prais & Winsten, 1954) method, based on the earlier, more primitive Cochrane–Orcutt method (Cochrane & Orcutt, 1949). The Prais–Winsten is a strong form of the more general Yule-Walker or "two-step full transform method" (Harvey, 1981). Also still used is a relatively primitive nonlinear least squares method, the Hildreth–Lu (Hildreth & Lu, 1960), which was improved on by the "nonlinear least squares" (NLS) method by Spitzer (1979). Comparative tests have concluded that the maximum likelihood ARIMA procedure (Box & Jenkins, 1976) is still the standard, and for small samples, the Prais-Winsten comes closest to that standard (Harvey, 1981; Harvey & McAvinchey, 1978; Judge, Griffiths, Hill, & Lee, 1985; Park & Mitchell, 1980). In this study, the best validated r_{auto} control method was used, the ARIMA AR1 (1, 0, 0) model.

A new analytic method such as Tau-U should show robustness to r_{auto}; that is, its magnitude and significance of results should not vary greatly under varying levels of r_{auto}. For SCR practitioners, an equally important standard of robustness is that r_{auto} should minimally distort graphed data. If removing or "cleansing" r_{auto} greatly distorts graphed data, it will prevent visual analysis, disallowing mutual validation by statistical and visual analysis. Cleansing data of r_{auto} should therefore minimally impact visual analysis. Most evaluations of robustness of statistical methods include the stability of standard error (SE) under various r_{auto} conditions. But that is not possible with Tau-U (or simple Tau), as its SE is based solely on the number of data points, which do not change under various levels of r_{auto}.

Despite its strengths, nonoverlap analysis is not best for some data series because it is insensitive to data trend. Trend is visually apparent in much graphed data, and is important to conclusion validity in two main ways. First, positive trend in the intervention phase is a valued measure of improvement not captured by mean shift or nonoverlap measures. Positive slope in the intervention phase suggests the likelihood of further improvement in the future, which is generally hoped for. Second, undesirable "preexisting" positive trend in the baseline phase suggests the client would have improved even without the intervention. Ignoring positive baseline trend risks erroneous conclusions about the cause of change. Current nonoverlap models cannot include baseline trend, as can the Allison et al. regression model (Allison & Gorman, 1993; Faith et al., 1997).

PROBLEMS IN BASELINE TREND CONTROL

Although the Allison et al. regression model (Allison & Gorman, 1993; Faith et al., 1997) does control for undesired baseline trend, unresolved issues still exist. The Allison et al. correction method involves semipartialling Phase A trend from the full original data set. But frequent users of the Allison et al. method encounter problems, some of which are demonstrated by an example (or "mis-example") data set, used throughout this paper.

Fig. 1a presents a short, simple AB design, with data points A: 2, 3, 5, 3 and B: 4, 5, 5, 7, 6. Means are A: 3.25 and B: 5.4. Regression slopes are A: .50 and B: .60. In Fig. 1a, the Phase A trend line has been extended through Phase B. This depicts the first step in the Allison et al. regression-based control (Allison & Gorman, 1993; Faith et al., 1997). Fig. 1b shows the transformed data following Phase A trend removal through semipartialling the prediction line from the original scores. These

Please cite this article as: Richard I. Parker, et al., Combining Nonoverlap and Trend for Single-Case Research: Tau-U, *Behavior Therapy* (2011), 10.1016/j.beth.2010.08.006

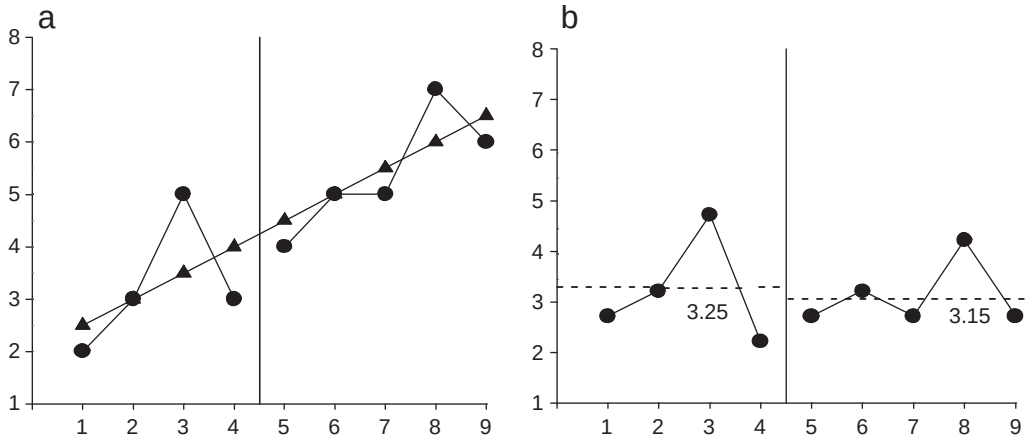


FIGURE 1 Example data set with (a) an illustration of control limitations, and (b) transformed data following Phase A trend control.

figures show that regression trend control is a powerful corrective. By controlling Phase A trend, the mean level of Phase B has been reduced below that of Phase A (Fig. 1b). Four concerns can be inferred from this example: (a) unreliability of Phase A trend, (b) no consideration of Phase A length, (c) questionable assumption that trend will continue, and (d) unintuitive mean comparisons after trend control. And a fifth problem not visible in this example is (e) no rational limits to change. Some of these interrelated problems have been previously identified (Scruggs & Mastropieri, 1998, 2001).

Unreliability of Phase A Trend

Extending Phase A trend into Phase B assumes a trustworthy Phase A trend line slope. That is because semipartialling Phase A trend is executed without regard to trend error. Most would visually judge that the Phase A trend in Fig. 1 is not pronounced or credible. In fact, its p value is only .49, and its slope has very wide confidence intervals (CI), spanning zero (85% CI is –.85, 1.85). Though lacking credibility, controlling it has considerable impact, both visually and on statistical results. The best solution to this dilemma appears to be carefully selective in when to apply baseline trend control. It should be applied only when Phase A trend is pronounced and statistically significant.

No Consideration of Phase A Length

Regression control of Phase A trend occurs without regard to Phase A length. Controlling Phase A trend from a phase of 5 or from 45 data points will have the same impact on Phase B data. Yet within a short baseline phase, the trend lacks credibility. A published SCR study may have a short baseline of 6 data points, followed by a longer intervention phase of 15 or 20 data points. In that case, Phase A trend control influence on Phase B data would seem excessive. A potential correction to this problem is to limit the application of baseline control to only longer A phases.

Questionable Assumption That Trend Will Continue

An assumption underlying trend control is that without intervention Phase A trend would continue unabated through Phase B. But that assumption may not be accurate. We examined a convenience sample of 160 published AB data sets, all of which had at least 10 data points in Phase A, to locate 34 with strong baseline within the first five data points trends (all at $p \leq 0.05$). To what extent did those strong trends continue through the next five data points (for a total of 10)? Thirteen of the 34 series (38%) were no longer significant at .05 and 10 were not significant at $p = .10$, even with the benefit of double the data points (10). In fact, the first five and second group of five data points in a data set bore little relationship to one another in trend. And what about those data sets where the first five data points lacked trend; did their next five data points show more trend? They did not, which suggests that the normal state of affairs for baselines is little or no trend, and measured baseline trend might be more apparent than real. Though far from conclusive, this finding raises doubts about the assumption of baseline trend and its routine control. To our knowledge this issue has not been formally studied.

Unintuitive Mean Comparisons after Baseline Trend Control

Results following trend control are rarely graphed, yet they should be. Visual analysts need access to data plots, and that includes the effects of baseline trend control. Figs. 1a and b show results of baseline trend control that are mildly problematic.

Please cite this article as: Richard I. Parker, et al., Combining Nonoverlap and Trend for Single-Case Research: Tau-U, *Behavior Therapy* (2011), 10.1016/j.beth.2010.08.006

Visual analysis of Fig. 1a indicates a rise in original data mean level, whereas Fig. 1b indicates a drop in mean level. To many, the conclusion of mean-level deterioration is not intuitive. The only present remedy may be to warn users that Phase A trend control transforms data to a point where visual analysis is no longer appropriate.

No Rational Limits to Change

The effects of baseline trend control also undesirably depend on Phase B length. Given longer B phases, Phase A trends tend to be projected outside the limits of the *y*-axis score scale, and resulting effect sizes will be unrealistically extreme. This predicament underscores the unbridled power of the control technique, which presently seems to be without a good solution. A partial remedy suggested has been to manually reset extreme predicted scores to within scale limits (Allison & Gorman, 1993). However, that sets artificial ceilings on effect sizes and violates the constant variance assumption.

An improved baseline trend control technique should have rational limits imposed on its impact. Rational limits could be based on the reliability of Phase A, Phase A length, and/or the length of Phase B. Baseline control presented in this paper within Tau-U does have rational limits on its impact.

COMBINING TREND AND NONOVERLAP

Mann-Whitney U

Common ground exists between data trend and nonoverlap in the nonparametric sampling distribution of "Kendall's *S*." The *S* distribution is the foundation for two established statistics: the Mann-Whitney U (MW-U) test of "dominance" or nonoverlap between two groups, and the Kendall Rank Correlation (KRC) coefficient. MW-U and KRC usually are employed to answer quite different research questions, and are applied to differently structured data sets. MW-U is an index of group (phase) difference in level (dominance), whereas KRC is a correlation index between paired score series. The MW-U computational algorithm first combines scores from two groups for a cross-group ranking. Those rankings are then separated and statistically compared for mean difference in ranks. This mean difference of ranks produces identical results to a pairwise comparison of all scores across groups (dominance). KRC uses the same algorithm for trend within a single group (Conover, 1999; Kendall & Gibbons, 1990), and produces identical results to MW-U if instead of two continuous variables (scores and time), the time variable is replaced by dummy codes (0 / 1) representing phases. The identity of MW-U and KRC permits nonoverlap and trend to be included within a single measure.

MW-U outputs two *U* values, larger (U_L) and smaller (U_S), of which the smaller is typically tabled in texts for inference testing. Their difference equals Kendall's *S* ($S = U_L - U_S$), which is the test statistic for significance of both MW-U and KRC (Hollander & Wolfe, 1999). Nonoverlap, or "percent of nonoverlapping data," can be calculated as the difference of the two *U* values divided by their sum: $(U_L - U_S) / (U_L + U_S)$; (Parker & Vannest, 2009). This formula can be simplified to: $S / (U_L + U_S)$, since $S = U_L - U_S$. The denominator, $U_L + U_S$ equals the total number of pairwise comparisons possible between two phases (two groups). That number is the product of the two group *N* ($n_1 \times n_2$), so for phases of 5 and 7 data points, the number of paired comparisons is $5 \times 7 = 35$. The MW-U nonoverlap statistic thus simplifies further to *S* / #pairs, which is literally "the proportion of pairwise comparisons that improve from Phase A to B," simplified to "the percent of nonoverlapping data between Phases A and B."

Kendall Rank Correlation

Kendall Rank Correlation (KRC) of two matched data series is presented in textbooks as quite different from MW-U, though their essential sameness is core to this paper. Underlying a KRC analysis on time and score is a simple algorithm. Scores are time ordered, and then all possible pairs of score data points are compared, in a "time-forward direction." Each pairwise comparison of scores is coded: (a) positive or improving over time (+), (b) negative or decreasing (–), or (c) tied (T). The total number of pairs is $N(N-1)/2$, where *N* equals the number of original scores. So a series of 8 scores has $(8 \times 7) / 2 = 28$ pairwise comparisons. *S* is calculated as the difference between the number of positive and negative codes: = #pos – #neg. Kendall's Tau equals *S* divided by the total number of pairs: = *S* / #pairs. For time–series data, Tau is therefore "the percent of all data pairs that show improvement over time," or more colloquially, "the percent of data that improve over time." Thus, for single-case research, KRC measures "trendedness" or the "tendency for scores to improve over time." Tau's direct interpretation is an asset over indices with more oblique interpretations such as Spearman Rho or least squares *R* or R^2 (Conover, 1999; Hollander & Wolfe, 1999; Sprent & Smeeton, 2007).

The "Pitman efficiency" (or power) of Kendall's Tau equals .91, the same as for Spearman Rho, so for well-conforming data, Pearson *R* requires a sample 91% the size of Tau to achieve the same power.

Please cite this article as: Richard I. Parker, et al., Combining Nonoverlap and Trend for Single-Case Research: Tau-U, *Behavior Therapy* (2011), 10.1016/j.beth.2010.08.006

When data do not meet parametric assumptions, then Tau can exceed Pearson *R* in power (to a Pitman efficiency of 1.27; Sprent & Smeeton, 2007).

MW-U and KRC Equivalence

From the foregoing, Tau trend and MW-U nonoverlap are the same. By formula, MW-U's "percent of nonoverlapping data" = $(U_L - U_S) / (U_L + U_S) = S / \#pairs = (\#pos - \#neg) / \#pairs$ = Tau. MW-U conducted on two groups and Tau conducted on a single time series are calculated the same way, have the same sampling distribution, and can be interpreted in the same manner. Percent of nonoverlapping or "improving" data between two phases is calculated the same as percent of improving data within a single phase. In both cases, all possible pairs of data are compared in a time-forward direction to obtain a net improvement sum, *S*. Both KRC and MW-U analyses can be interpreted as nonoverlap or as trendedness. This manuscript emphasizes the trendedness interpretation for both KRC and MW-U.

KRC calculates not linear trend, but rather monotonic trendedness, or the tendency for scores to improve over time, following any profile or configuration (Conover, 1999; Hollander and Wolfe, 1999; Sprent & Smeeton, 2007). Monotonic trend does not assume that a straight line will be a good summary of the path of improvement. So Tau reflects both monotonic trend and the percent of data that improve over time—they are the same. And both of these can also interpret trend between phases as "percent of data that improve in a time forward (from Phase A to B) direction," which is also "percent of nonoverlapping data." The one computational difference between MW-U and KRC is that KRC stipulates a single *N* (number of data pairs), whereas MW-U requires an *N* for each phase (n_1 and n_2) which affects calculation only of the variance and standard error.

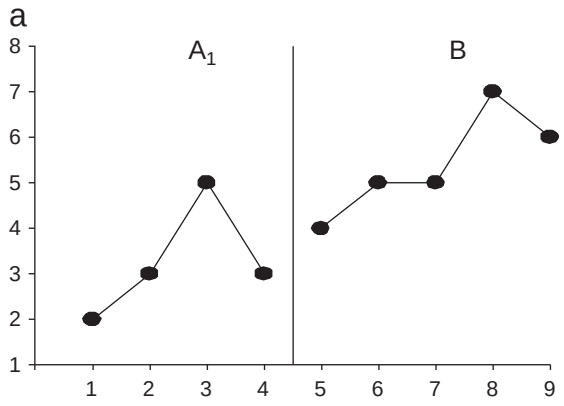
KRC and MW-U Inference Tests

Both KRC and MW-U rely on the *S* distribution for significance testing; "a test of Tau is a test of *U*" (Armitage et al., 2002, p. 279). For smaller samples of $N < 10$, both KRC and MW-U should use an exact permutation test, which is commonly offered in statistical software packages. Exact inference tables for $N < 10$ are also available in nonparametric and biostatistics textbooks. For $N \geq 10$, the *S* distribution rapidly approaches normal, so the test statistic $z = S / SE_S$ can be used for both KRC and MW-U. From KRC, SE_S is usually output directly. From MW-U, only SE_{rank} is output directly, and $SE_S = 2 \times SE_{rank}$.

Many KRC and MW-U modules provide full significance test output: SE_S, *z* scores, and exact permutation *p* values. An accurate SE_S for nonoverlap between two phases can be obtained from either a MW-U or KRC module. This paper uses a KRC module for all analyses, because only KRC can also measure within-phase trend. To test an A versus B phase shift by a KRC module, two variables are entered: scores, and a categorical phase variable that is "dummy coded" (0 / 1) or by a mixed code (explained later). The output from KRC for *S* and SE_S will be accurate, and will match output from an MW-U module. *Note*: The KRC output for Tau will not be accurate because of the use of the dummy code, so Tau must be calculated by hand. The name given to this new analysis merging trend and nonoverlapping data is "Tau-U," after its parents: Kendall's Tau and Mann-Whitney U.

Example AB Design Data

Tau-U is best described by application to sample data. Fig. 2a is the same short AB design graph (A: 2, 3, 5, 3; B: 4, 5, 5, 7, 6) from Fig. 1a. Beside it,



b

		B					A			
		6	7	5	5	4	3	5	3	2
	2	+	+	+	+	+	+	+	+	0
	3	+	+	+	+	+	T	+	0	
A	5	+	+	T	T	–	–	0		
	3	+	+	+	+	+	0			
	4	+	+	+	+	0				
	5	+	+	T	0					
B	5	+	+	0						
	7	–	0							
	6	0								

FIGURE 2 Example time series data with (a) A and B phase, (b) difference matrix of Fig. 2a data with all pairwise data comparisons, made in a "time-forward" direction. The rectangular box in the center represents between-phase data, and the two adjacent rectangular areas represent within Phase A or B.

Please cite this article as: Richard I. Parker, et al., Combining Nonoverlap and Trend for Single-Case Research: Tau-U, *Behavior Therapy* (2011), 10.1016/j.beth.2010.08.006

Fig. 2b is a difference matrix of all pairwise data comparisons made in a "time-forward" direction. The left margin of Fig. 2b contains the data series, and atop the matrix is the same series, in reverse order. A matrix like this is used to explain Kendall's Tau in biostatistics and nonparametric textbooks. But the difference in this figure is that the data have been segmented or partitioned between Phases A and B, to distinguish the pairwise comparisons that contribute to the A versus B contrast (gray-shaded rectangle) from those that contribute to within-phase trend (the two triangles).

The Fig. 2b matrix contains "+" at the intersect of each data pair for which the later value is larger, and "–" when the later value is smaller. Ties, denoted *T*, are not analyzed in this paper, but would be included to calculate a variation of Tau, Tau-b. Tau was chosen for this paper over Tau-b for three reasons: (a) only Tau offers exact permutation tests, (b) Tau is simpler to compute, and (c) Tau is more conservative. The difference between Tau and Tau-b tends to be minimal unless there are many ties, which will inflate Tau-b (Armitage et al., 2002). Both Tau and Tau-b are well-respected indices.

Fig. 2b shows the full matrix of $N(N-1)/2 = (9\times8)/2 = 36$ pairs within three partitions. The figure includes the A versus B nonoverlap contrast (rectangle in upper left), trend within Phase A (upper-right triangle), and trend within Phase B (lower-left triangle). These three components comprise all sources of trend in the full series of 9 data points. By considering only selected components, we may draw conclusions about intervention effectiveness. From the rectangle alone, phase nonoverlap can be calculated. The rectangle and lower triangle (Phase B) together summarize two valued outcomes: phase nonoverlap and Phase B improvement trend. Subtracting the upper triangle (Phase A) from the rectangle gives nonoverlap with Phase A trend controlled. Subtracting the upper triangle from the combined rectangle and lower triangle summarizes nonoverlap plus Phase B trend, after control of Phase A trend.

The Fig. 2b matrix strengthens the rationale for mixing phase nonoverlap with monotonic trend. Tau for each of the three matrix components can be summarized by *S* / #pairs, with similarly computed SE_S. The three components can be added and subtracted via *S*, weighted by number of paired comparisons (#pairs). These #pairs can be counted in Fig. 2b: 20 for phase nonoverlap, 6 for A trend, and 10 for B trend. Thus, the A versus B contrast is weighted more heavily than the two within-phase trends when considering overall trendedness of the data. The Fig. 2b matrix is presented as a logic model; it is not needed for Tau-U calculations.

The A versus B rectangle in the upper-left corner of Fig. 2b contains results of $n_A \times n_B = 4\times5 = 20$ paired comparisons. For this contrast, $S = (\#pos - \#neg) = 17 - 1 = 16$, and $Tau = S / \#pairs = 16/20 = .80$. This phase contrast can be analyzed alone, in a MW-U module, yielding $U_L = 18$, $U_S = 2$, $S = 16$, $SE_S = (2\times SE_{rank}) = (2\times3.996) = 7.99$, $z = (S / SE_S) = 2.01$, and two-tailed $p = .045$. The MW-U module does not provide Tau, but it is calculated as $(U_L - U_S) / (U_L - U_S) = (18-2)/(18+2) = .80$. Identical results are obtained from analyzing the same contrast in a KRC module (StatsDirect was used, with phase coded 0 / 1): $S = 16$, $SE_S = 7.99$, $z = 16/7.99 = 2.00$, two-tailed $p = .045$, and the exact permutation (for $N<10$) two-sided $p = .119$. *Note*: The Tau value output from this KRC analysis (with phase coded 0 / 1) will not be accurate, so it must be calculated as $S / \#pairs = 16/20 = .80$. The remaining two triangular partitions of Fig. 1b represent trends within Phase A (upper right) and Phase B (lower left). Each can be analyzed separately within KRC to confirm *S* and Tau values. For the Phase A triangle: $S = (4-1) = 3$, $Tau = S/\#pairs = 3/6 = .50$, $SE_S = 2.769$, $z = 1.08$, two-sided $p = .28$, and exact permutation $p = .33$. For the Phase B triangle: $S = (8-1) = 7$, $Tau = S / \#pairs = 7/10 = .70$; $SE_S = 3.96$, $z = 1.77$, two-tailed $p = .007$, and exact $p = .08$.

Finally, demonstrating the additivity of the matrix components, a single traditional KRC analysis conducted on the full data series of $N = 9$ (time and scores input) is included in Table 1. The results are $\#pairs = 36$, $\#pos = 29$, $\#neg = 3$, $S = 26$, $Tau = 26/36 = .722$, $SE_S = 9.345$, approximate $z = 2.78$, $p = .008$, exact $p = .006$. Table 1 contains six data columns, all with computer output data (StatsDirect). The first three data columns are for the three partitions of the matrix, and the fourth column pertains to the full data series. Partitioning the matrix is analogous to partitioning an ordinary least squares variance matrix. Across the first three data columns, the values #pairs, #pos, #neg, and *S* are strictly additive. Tau values are additive after proper weighting by their respective #pairs. SD_S are not additive, but their squares, VAR_S, are practically additive. The sum of the first three $VAR_S = 63.89 + 7.67 + 15.67 = 87.23$, and $(87.23)^{1/2} = 9.34$, which equals the SD_S value output by a KRC module for the full series. The final two data columns are described later.

INTERPRETATION OF TAU-U RESULTS

Tau-U is actually a family of four indices, three of which include nonoverlap with trend together: (a) A versus B phase nonoverlap, (b) nonoverlap and

Please cite this article as: Richard I. Parker, et al., Combining Nonoverlap and Trend for Single-Case Research: Tau-U, *Behavior Therapy* (2011), 10.1016/j.beth.2010.08.006

Table 1
Example Tau-U Analysis

	Partitions of Matrix			Full Data Matrix	Tau-U Analysis	
	A vs. B	$Trend_A$	$Trend_B$		A vs. B + $trend_B$	A vs. B + $trend_B$ – $trend_A$
#pairs	20	6	10	36	30	36
#pos	17	4	8	29	25	26
#neg	1	1	1	3	3	6
S	16	3	7	26	23	20
Tau	16 / 20 = .80	3 / 6 = .50	7 / 10 = . 70	26 / 36 = .72	23 / 30 = .77	20 / 36 = .56
SD_S	7.99	2.79	3.96	9.35	8.91	9.35
VAR_S	63.89	7.67	15.67	87.33	87.22	87.33
Z	2.00	1.08	1.77	2.78	2.58	2.14
p (*Z* based)	.05	.28	.007	.008	.0098	.032
p (exact)	.12	.33	.08	.006	.0127	.045

Phase B trend together, (c) nonoverlap with baseline trend controlled, and (d) nonoverlap and Phase B trend with baseline trend controlled. The first of these, A versus B, is very similar to the nonoverlap of all pairs (NAP; Parker & Vannest, 2009). This paper emphasizes that the A versus B results may be interpreted either as nonoverlap: "percent of nonoverlap between phases," or as trendedness: "percent of data showing improvement between phases." The second summary, nonoverlap and Phase B trend together, is "percent of data showing improvement between A and B, and within Phase B." It is analogous in regression to predicting scores from both phases and a dummy-coded time variable with the Phase A portion filled with the Phase A mean, and the Phase B portion filled with the time values: (A: 3.3, 3.3, 3.3, 3.3; B: 5, 6, 7, 8, 9). There is an important difference in how trend behaves in Tau-U versus regression analysis. In regression, including time as a predictor with phase (with a Time × Phase interaction) always equals or improves on results from phase as the sole predictor. But in Tau-U, including a Phase B trend with NAP nonoverlap can easily reduce results. That is because in the Tau-U additive model, by including Phase B trend, one also includes additional variance beyond that in the phase nonoverlap contrast only.

The third summary, "nonoverlap with baseline trend controlled," is most related to the Allison regression control method (Allison & Gorman, 1993; Faith et al., 1997) by partialling Phase A trend out of the entire data series. The Allison method results in zero Phase A trend, so the final analysis tests only the mean in Phase A (and a reduced trend in Phase B). But the Tau-U results need to be interpreted differently, due to the different control method. Regression trend control is via a vector, whereas Tau-U controls for only a fixed amount of trendedness, limited by the length (#pairs) of Phase A. Tau-U trend control thus has a smaller impact on results, which is considered an advantage, given concerns expressed earlier about baseline trend over control. Compared to regression control by vector, the Tau-U subtraction is constrained by amount of Phase A trend, by Phase A length, and by the relative lengths of Phase A and Phase B; therefore, Tau-U does not control baseline trendedness beyond rational *y*-scale limits. But the final summaries from regression and Tau-U are defined similarly.

The fourth Tau-U summary, "nonoverlap with Phase B trend with baseline trend controlled," simply adds to the third model the weighted Phase B trend. As with the second and third models, adding within-phase trend also adds variance from a new partition in the agreement matrix, so it can increase or reduce results from a simpler model. Adding Phase B trend to the A versus B contrast may increase or decrease effect size results. The analog to this fourth Tau-U is the Allison baseline correction technique, the final step of which is an MTS (Mean × Trend Shift) regression analysis (Allison & Gorman, 1993; Faith et al., 1997). The Allison MTS R^2 can be interpreted as "the proportion of variance accounted for by AB shift and B trend, after control of Phase A trend." The Tau-U summary index is interpreted as "the percent of data that improve over time considering both phase nonoverlap and Phase B trend, after control of Phase A trend."

Answering Questions About Improvement

Tau-U, the index of between and within-phase trend, is useful for answering at least four research questions in SCR. The first two presented below require only simple KRC or MW-U analyses, so are not new or novel. The novel Tau-U is needed to answer the third and fourth questions, which combine within-phase monotonic trend and AB

Please cite this article as: Richard I. Parker, et al., Combining Nonoverlap and Trend for Single-Case Research: Tau-U, *Behavior Therapy* (2011), 10.1016/j.beth.2010.08.006

nonoverlap in a single improvement index. Each question is followed by data input procedure, output, and the solution.

1. What is the improvement trend during Phase B?
   a. *Input*: To KRC, variables score and time for Phase B only.
   b. *Output*: The improvement trend (Tau) should be output. If not, calculate Tau as *S* / #pairs, where #pairs = n(n − 1) / 2 = 5(5 − 1) / 2 = 10.
   c. *Solution*: Here, Tau = 7 / 10 = .70, so 70% of the intervention phase data show improvement, and this improvement trend is borderline significant (exact p = .08).
2. What is the improvement in nonoverlapping data between Phase A and B? (KRC directions are given here. The same results are also obtainable from MW-U.)
   a. *Input*: To KRC, variables score and phase (coded 0 / 1).
   b. *Output*: Collect *S* = 16, and calculate #pairs as $n_A \times n_B = 4 \times 5 = 20$. Calculate Tau = *S* / #pairs = 16 / 20 = .80. The SD_S (7.99), z (2.00), and p values from KRC are output accurately (but neither Tau nor #ties will be accurate).
   c. *Solution*: From Phase A to B, data show an 80% improvement trend (or 80% nonoverlap), which is statistically significant at $p < .05$ from a normal distribution approximation, but at only p = .12 from an exact permutation test.
3. What is the overall client improvement in A versus B contrast plus Phase B trend?
   a. *Input*: To KRC, score and a modified time variable composed of zeros for Phase A, and the normal time sequence for Phase B: (0, 0, 0, 0|5, 6, 7, 8, 9).
   b. *Output*: Obtain *S* = (25 − 2) = 23. Add #pairs for A versus B (4 × 5 = 20) to #pairs for B (5 × 4) / 2 = 10 to obtain total #pairs = 30. Calculate Tau = *S* / #pairs = 23 / 30 = .77. As output from KRC, the SD_S (8.908), z (2.581), and p values are accurate.
   c. *Solution*: Data showed 77% overall improvement between phases and during treatment. This amount of improvement is significant at p = .0098, or at p = .0127, using an exact inference test.
4. What is the overall client improvement, controlling for preexisting (baseline) improvement trend? Phase A trend can be "controlled" through the entire data series by reversing its sign, and then recomputing the full trend. (*Note*: Reversing signs affects only *S* and Tau, not SD_S, z, or p.) This technique imposes a rational maximum or ceiling on control (unlike OLS regression analysis). The trend reduction cannot exceed Phase A trend's negative value. There are multiple ways to do this Tau-U analysis, all with the same result (see Table 1).Two are presented here.

   Control Method 1:
   a. *Input*: In the time variable, backward-code Phase A: 4, 3, 2, 1. Maintain the true time values for Phase B: 5, 6, 7, 8, 9. Conduct a KRC analysis.
   b. *Output*: All program output will be accurate: #pos (26), #neg (6), *S* (20), SD_S (9.345), z (2.14), and inference tests. The Tau value will be accurate, but may be the Tau-b version, depending on the software used. So it is best to calculate your own Tau = *S* / #pairs = 20 / 36 = .56.
   c. *Solution*: Controlling for phase A improvement trend, overall improvement (in both A *vs* B and within-phase B trend) is reduced to 56%, with, approximate p = .03, and exact p = .045.

   Control Method 2: Replace the Phase A *S* value (+3) with its negative (−3). Then recalculate Tau for the full matrix as *S* / #pairs = (16 + 7 − 3) / 36 = .56. The SE_S does not change from that of the full model, so output is still $z = S / SE_S = 20 / 9.35 = 2.14$.

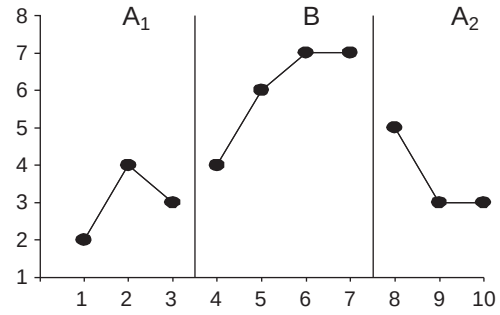
A Second Example

The second example is an ABA reversal design of 10 data points total, made short to permit easy replication. Figs. 3a and 3b show the graph and its Tau matrix. The matrix includes six partitions: three phase trends (A_1, B, A_2) and three phase contrasts (A_1 vs. B, B vs. A_2, A_1 vs. A_2), of which the last, A_1 versus A_2, is not relevant. Note that the matrix is not essential to calculations, and is included here only as a heuristic.

For this second data set, only the second and third questions are answered, and more briefly:

1. What is the improvement between phases? This question implies both B versus A_1 and B versus A_2 contrasts. In SCR, contrasts of adjacent phases are usually defensible, but between separated phases often are not.
   a. *Input*: To KRC, scores and time. Time is coded 0, 0, 0|1, 1, 1, 1|0, 0, 0.
   b. *Output*: Collect *S* (21), and calculate #pairs as $(N_{A1} \times N_B) + (N_B \times N_{A2}) = 24$.

a



b

		A2			B				A1		
		3	3	5	7	7	6	4	3	4	2
	2	+	+	+	+	+	+	+	+	+	0
A1	4	–	–	+	+	+	+	T	–	0	
	3	T	T	+	+	+	+	+	0		
	4	–	–	+	+	+	+	0			
B	6	–	–	–	+	+	0				
	7	–	–	–	T	0					
	7	–	–	–	0						
	5	–	–	0							
A2	3	T	0								
	3	0									

FIGURE 3 Example time series data with (a) A_1, B, and A_2 phases (b) difference matrix of Fig. 3a data with all pairwise data comparisons.

The analysis does not contrast Phases N_{A1} and N_{A2}. Tau = *S* / #pairs = 21 / 24 = .88. Output for SD_S (9.21), *z* (2.28), and *p* will be accurate.

c. *Solution*: Phase B contrasted with A_1 and A_2 shows over 87% improvement, significant at $p = .02$ (exact $p = .07$).

2. What is the overall improvement, considering phase contrasts plus growth within the intervention phase?
   a. *Input*: To KRC, scores and phase, coded 0, 0, 0|4, 5, 6, 7|0, 0, 0.
   b. *Output*: Collect *S* (26). Total #pairs expands from the first analysis to include within-Phase B trend: (4 × 3) / 2 = 6. So #pairs = 12 + 12 + 6 = 30. Calculate Tau = *S* / #pairs = 26 / 30 = .87. Output for SD_S (9.63), *z* (2.70), and *p* will be accurate.
   c. *Solution*: Overall improvement trend (between phases plus within Phase B) equals 87%, which is significant at approximate $p = .007$, or by exact test, $p = .017$. Note that this is the same Tau calculated immediately above, but with a stronger *p* value. The 87% Tau did not change with the addition of Phase B trend because it existed at the same level in both B trend and AB contrast. Our gain here by including Phase B trend is in greater statistical power; the active *N* in the analysis increases, and with it the number of comparisons, yielding a more favorable *p* value.
3. What is the overall improvement, considering phase contrasts and intervention phase trend, and also controlling for initial baseline trend? (*Note*: In reality this baseline trend is not pronounced or reliable so we would find its control difficult to justify in real life. It is controlled here only to demonstrate the procedure.) There are also multiple ways to conduct this analysis, but only one method is demonstrated here.
   a. *Input*: First run the KRC, as above on data coded 0, 0, 0|4, 5, 6, 7|0, 0, 0, which results in Tau = *S* / #pairs = 26 / 30 = .87.
   b. *Output*: Next obtain the *S* value for Phase A only: *S* = 2 – 1 = 1. Change its sign to negative, and combine with the previous result: Tau = (26 – 1) / 30 = 25 / 30 = .83.
   c. *Solution*: Overall improvement is 83%, including two phase shifts, Phase B improvement, and controlling for Phase A trend.

FIELD TESTING THE TAU-U

The purpose of a field test is to give potential users a sense of how Tau-U performs with typical data, particularly how much of a change is caused by including Phase B trend with AB nonoverlap, and also by the optional Phase A trend control. These two features mark the difference between the new Tau-U and the two simpler indices: Tau trend (from KRC), and phase nonoverlap (from MW-U), which is quite similar to NAP (Parker & Vannest, 2009). Tau nonoverlap scores correlated at Rho = .92 with regression (or *t*-test based) R^2 effect sizes, at Rho = .76 to .93 with other effect sizes, and at Rho = .84 with visual judgments of client improvement (Parker & Vannest, 2009). Tau-U is new, and readers need to know whether and how much those results change by adding or controlling for trend.

Field testing consisted of applying Tau-U to 382 simple AB contrasts from published articles. The graphs had been digitized in stages over recent years, from articles published in the past 25 years. This was a convenience sample, including all articles that had clearly digitizable graphs, without regard for design type, target behavior, or intervention. Articles included a mix of academic and behavioral outcomes. Leading journals in special education, school psychology, and behavioral psychology were well represented in this convenience sample. For complex, multiphase designs, only the initial A and B phases were included.

Please cite this article as: Richard I. Parker, et al., Combining Nonoverlap and Trend for Single-Case Research: Tau-U, *Behavior Therapy* (2011), 10.1016/j.beth.2010.08.006

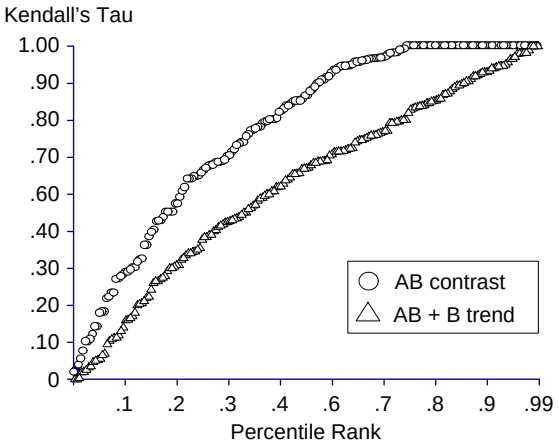


FIGURE 4 Probability plot comparing 176 data sets with Phase B trends and AB contrasts in the same direction.

Details of the collection and digitizing have been previously reported (Parker et al., 2005).

Questions that potential users of Tau-U would likely have include (1) What is the impact of adding Phase B trend to nonoverlap? (2) What are the distribution characteristics of a Tau-U (overlap plus B trend) index? (3) What is the need for controlling baseline trend, and what is the impact? and (4) How does Tau-U respond with autocorrelated data?

Question 1: What is the impact of adding Phase B trend to nonoverlap? The influence of Phase B trend on an A versus B (AB) contrast can be calculated by the weight of the Phase B trend (#pairs) relative to the weight of the A versus B contrast. Suppose $n_A = 8$ and $n_B = 6$, so the AB contrast has a weight of $8 \times 6 = 48$ pairs. The weight for Phase B trend only, calculated from its $n_B(n_B - 1) / 2$ pairs, equals $6 \times 5 / 2 = 15$. So Phase B trend contributes 15 / 15 + 48 = 24% of the final Tau-U. Suppose the AB contrast yields Tau = .50, and the Phase B Tau = .60. Tau-U for these two sources of improvement together (AB + B trend) will be $.50 \times 76\% + .60 \times 24\% = .52$. Though Phase B has a stronger trend, its influence is limited by its fewer paired comparisons than in the AB contrast.

The field test showed within-phase trends to be few and weak, compared to nonoverlap magnitudes from AB contrasts. Only 176 of the 382 data sets (46%) had AB contrasts in the same direction as their Phase B trend. Of these 176, including B trend with the AB contrast caused a smaller Tau-U index in 74% (130 data sets). Tau-U increased due to adding Phase B trend in only 26% (46 data sets). In those 130 data sets where trend decreased, it decreased by 15%. In the 46 where it increased, it did so by a larger 56%, on average.

A second impact of including Phase B trend was on significance levels. Adding Phase B to an AB contrast increased the number of paired comparisons (#pairs) by 23%, on average. This improved p values by an average of .02 to .05, depending on the overall N and the ratio of n_A to n_B. Results previously not significant at $p = .10$ gained significance at $p = .05$. The improvement was greatest for the shorter data sets.

Question 2: What are the distribution characteristics of a Tau-U (overlap plus B trend) index? Tau-U's usefulness depends partly on its ability to discriminate among results from different studies. Given a large sample, a "uniform probability plot" can indicate discriminability (Cleveland, 1985). Strong discriminability is seen as a diagonal line, without floor or ceiling effects, and without gaps, clumping, or flat segments (Chambers, Cleveland, Kleiner, & Tukey, 1983; Hintze, 2006). Fig. 4 contains a probability plot comparing the 176 data sets with Phase B trends and AB contrasts in the same direction.

Tau-U presents a nearly ideal profile, compared to the simpler AB nonoverlap, which has a pronounced ceiling around Tau-U's 75th percentile. That ceiling is a shortcoming of all nonoverlap techniques; beyond complete nonoverlap, effect sizes cannot increase. Tau-U shows no ceiling or floor effects, gaps, or clumping. Noteworthy in Fig. 4 are the generally higher scores from the simple AB contrast. Considering that only nonoverlap gives larger results, being sensitive also to phase trend typically gives more modest results. The differences between AB contrasts and Tau-U appear large on the graph; .10 to .20 points over much of the distribution.

Table 2 gives quartile markers for the same results ($N = 176$) displayed in Fig. 4. The first two rows of the table are calculated from actual values and the last two from absolute values.

Table 2 confirms that the AB contrast hits a ceiling around its 75th percentile. It also confirms the score spread of nearly .1 to .2 points between the AB contrast and Tau-U for the middle half of the scores. These smaller scores are closer to R and R^2 scores for the same data. At each of the five

Table 2
Quartile Markers for AB Contrast and Tau-U (with B Trend)

Analysis	Quartile				
	10th	25th	50th	75th	90th
AB contrast	–.997	–.87	.96	.92	1.00
Tau-U	–.80	–.62	.16	.73	.91
abs AB contrast	.26	.60	.88	1.0	1.00
abs Tau-U	.29	.48	.66	.82	.93

Please cite this article as: Richard I. Parker, et al., Combining Nonoverlap and Trend for Single-Case Research: Tau-U, *Behavior Therapy* (2011), 10.1016/j.beth.2010.08.006

quartile markers, Pearson *R* always fell between the AB contrast and Tau-U.

Question 3: What is the need for controlling baseline trend, and what is the impact? A trend level of .40 or 40% was selected ad hoc as a level high enough to be of interest in most data sets. Tau = .40 represents the 75th percentile for the published Phase A trends, that is, 25% of the data had trends at ± .40 or more extreme. Only those data sets with Tau ≥ .40 in both Phase A and in the AB contrast (and both trends in the same direction) were selected for baseline trend control. That selective criteria resulted in only 31 candidates for Phase A trend control.

Removing baseline trends had the effect of reducing the simpler (AB + B_{trend}) values from a median of .91 to .62, a reduction of .29 Tau points, which is 32% median reduction from the original AB + B_{trend} value. The IQR range around that 32% was 10 to 55%. Considering that these 31 data sets included the most extreme positive Phase A trends, a 10 to 55% change is not large. These results were compared with the impact of Allison et al. regression control (Allison & Gorman, 1993; Faith et al., 1997). The same 31 data sets underwent Phase A regression (semipartialling) control. The regression control nearly cut in half the AB + B_{trend} results, the median R^2 dropping from .74 to .38 or 48% reduction. The actual difference in effect size reduction by regression control (48%) and Tau-U control (36%) is likely even greater than obtained. The Tau-based selection of these 31 data sets maximized the likelihood for Tau change, not for R^2 change. Had R^2 selection criteria been used, the regression control would show relatively greater change, and Tau-U relatively less change.

Question 4: How does Tau-U respond with autocorrelated data? A statistical method is robust to r_{auto} if its magnitude and its significance do not vary greatly with small and medium levels of positive r_{auto}. Robustness to r_{auto} is often ascertained by Monte Carlo studies, but those studies are problematic in SCR, where 100 studies may be represented by almost as many different scales, both interval and ordinal, both categorical and continuous, some with little central tendency, and all varying in upper and lower limits. Simulating that scale range may not be practicable. Therefore, this study evaluated Tau-U robustness to r_{auto} by checking it individually on the 365 published AB data sets. r_{auto} was checked in data sets before and after they had been cleansed of r_{auto} by the best established method, the ARIMA Lag-1 (1, 0, 0) model. The primary criterion for robustness to r_{auto} was minimal change in the Tau-U result from before to after cleansing. As noted earlier, another criterion, impact on standard error, cannot be applied to Tau-U. A second criterion that was included, but considered only informally, was the extent of distortion of graphed data due to r_{auto} removal.

r_{auto} cleansing was applied to only those data sets showing positive levels > +.20. Of the total 367 data sets, 151 (41%) showed large (>.20) negative r_{auto}, 86 (23%) showed small (<.20) r_{auto}, 58 (16%) showed small positive r_{auto}, and 72 (20%) had large (>.20) r_{auto} that needed cleansing.

The 72 data sets were cleansed via an iterative ARIMA maximum likelihood analysis, employing a Lag-1 autocorrelation model for each phase separately, after detrending each phase for linear growth. The ARIMA cleansing was largely successful, as seen by comparing the distribution of r_{auto} percentiles on (original, cleansed) scores: 10th (.23, –.06), 25th (.30, .00), 50th (.44, .06), 75th (.64, .11), and 90th (.71, .19). Of the 72 data sets for which r_{auto} cleansing was attempted, it was not wholly successful with only six, which remained with r_{auto} above +.20. Those six data sets all began with high, positive levels of r_{auto}, five of them at r_{auto} –.57 or above.

The question of change in Tau-U values is addressed by Table 3. Table 3 shows percentile distributions of Tau-U values before and after cleansing, along with the amount and percent of change (percent of original Tau-U).

There was no systematic direction of Tau-U change from before to after cleansing r_{auto}. And for approximately 75% of the data sets, changes in Tau-U would be considered minor. However, for 25% of the cleansed data series, Tau-U value changes were substantial.

Fig. 5 illustrates for informal scrutiny typical changes in graph configuration from original to cleansed data. The four graphs represent successful removal of different initial levels of r_{auto}. Original scores are circles, and cleansed scores are triangles. Fig. 5a shows initially low r_{auto} of .21 reduced to .01 after cleansing, but Tau-U changed .65 to .56. In Fig. 5b, r_{auto} was controlled from .34 to –.04,

Table 3
Percentile Distributions of Tau-U with Autocorrelation Cleansing

	Percentile				
	10th	25th	50th	75th	90th
Original Tau-U	.39	.72	.95	1.00	1.00
Cleansed Tau-U	.34	.75	.95	.99	1.00
Change amount	.00	.00	.01	.04	.14
Change percent	.00	.00	.01	.07	.46

Please cite this article as: Richard I. Parker, et al., Combining Nonoverlap and Trend for Single-Case Research: Tau-U, *Behavior Therapy* (2011), 10.1016/j.beth.2010.08.006

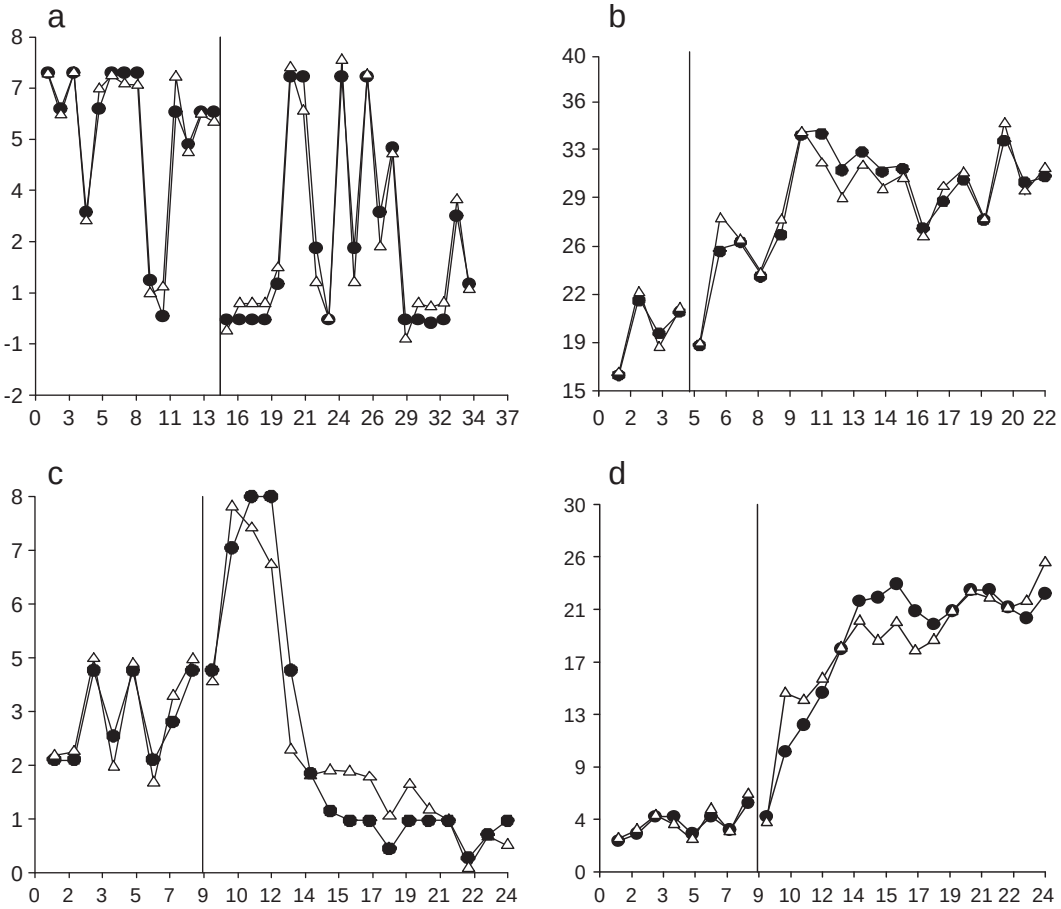


FIGURE 5 Shows four example data sets with varying degrees of autocorrelation. This figure also illustrates the amount of distortion that occurs in data when (a) low $r_{auto,}$ (b) medium-low r_{auto}, (c) high-medium r_{auto}, and (d) high r_{auto} is cleansed.

and NAP changed little, from .92 to .94. In Fig. 5c, r_{auto} of .51 dropped to .08, and Tau-U changed only from .48 to .42. Finally, Fig. 5d shows high r_{auto} of .64 eliminated to –.01, and Tau-U changed very little, from .96 to .95.

Readers can judge whether the graph distortion due to cleansing is tolerable. In general, the greater the r_{auto} cleansed, the greater the graph distortion. However, changes were often restricted to certain graph segments, as in Fig. 5d. In Fig. 5d, the change is nearly all in Phase B, with Phase A change barely noticeable.

Discussion

This paper presented Tau-U, a family of indices that can combine Phase AB nonoverlap with Phase B trend, and that permit control of undesirable positive Phase A trend. Tau-U was presented as an alternative to both regression-based models and to simpler dominance-based (nonoverlap) models. It was demonstrated that nonoverlap between phases and trend within phases can both be calculated from a single statistic, KRC, and both with an *S* sampling distribution. It was demonstrated and documented by expert sources that the KRC trend test and MW-U test between groups are statistically the same.

Tau-U was presented in the context of a rapidly developing field of statistical analysis for single-case research. The two existing analytic models of regression and simple nonoverlap or dominance were both shown to have weaknesses. Regression, the most comprehensive and flexible of the two, violates data and scale-type assumptions more often than not. Nonoverlap models lack statistical power, do not discriminate well among the more successful interventions, and cannot give credit for improvement trend during an intervention. A final problem with regression was how it controls positive baseline trend (through semipartial correlation). That method of control was argued to (a) produce extreme results, (b) not attend to measurement error of the Phase A trend, (c) yield sometimes nonsensical results, and (d) rely on a questionable assumption of continuing trend.

Please cite this article as: Richard I. Parker, et al., Combining Nonoverlap and Trend for Single-Case Research: Tau-U, *Behavior Therapy* (2011), 10.1016/j.beth.2010.08.006

Tau-U can potentially address the limitations of both regression and of simple AB nonoverlap. Like regression, it is a complete measure, including both trend and level. In addition, it is distribution free, and controls positive baseline trend in a more defensible manner than does the regression-based Allison et al. approach (Allison & Gorman, 1993; Faith et al., 1997). Tau-U is analogous to the Allison et al. regression model where Phase A trend has been controlled and both AB mean shift and the remaining Phase B trend contribute to the final R^2. But trendedness in Tau-U is dissimilar to regression slope; it is more closely related to R or R^2. Similar to R^2, Tau-U's trendedness is the percent of data that improve over time, but monotonically, in any profile—not only in a straight line.

Like other nonoverlap techniques, Tau-U is "distribution free," with minimal data assumptions. But Tau-U containing both AB nonoverlap and Phase B trend is unlikely to hit a 100% ceiling, which is not the case with other simpler nonoverlap techniques. This characteristic gave Tau-U superior discriminating power among our sample of published data series, compared to a simple AB nonoverlap analysis that could not discriminate among the top quarter of results.

Inclusion of Phase B trend typically decreased rather than increased Tau-U. By adding Phase B trend we also add additional variance (#pairs). The weighted S for A versus B nonoverlap tended to be larger than the weighted S for within-phase trend. A small improvement trend of 30% in Phase B combined with a typically larger, for example, 90% nonoverlap, will result in a Tau-U between these two figures, though closer to the 90%. Also, the negative impact of a weak Phase B improvement trend is limited by the proportional length of Phase B. For two phases of 5 data points each, negative impact of a very small positive B trend is limited to its proportional weight (#pairs), which is 25 pairs for the AB contrast, and only 10 pairs for Phase B trend.

Likewise, controlling trend from Phase A is conservative and measured. Unlike regression, baseline trend is not a vector that is assumed to continue ad infinitum. The impact of removing trend is limited by the Phase A length, that is, its number of paired comparisons. In most SCR studies the interventionist anticipates both a level shift/jump in performance and an improvement trend into the future. With a simple mean shift, median shift, or nonoverlap index, only part of that expectation is being measured. Failure to measure Phase B improvement trend in the effect size risks losing focus on improvement rate.

Tau-U was only somewhat influenced by autocorrelation (R_{auto}). Tau-U magnitude and graph configurations were monitored, but only the first was formally examined. For 75% of the data sets with dangerous levels of autocorrelation, its removal changed Tau-U values little. But for the remaining 25% of the 72 data sets, changes were larger. Thus, although Tau-U is "distribution free," it is not impervious to Lag-1 autocorrelation. But to keep perspective on the problem, the Tau-U values that showed substantial change from removing dangerous levels of R_{auto} numbered only 18 out of 367, or less than 5% of the original sample. R_{auto} does not impact Tau-U's standard error (and significance level), as its SE is based solely on the number of data points per phase.

There are cases where Tau-U with B trend need not be used. If a positive trend is impossible, quite unlikely, uninteresting, or undesirable, then B trend need not be included in an effect size. Also, if Phase B trend is impossible because performance has hit a scale ceiling, then Tau-B should not be used. Otherwise, for those many cases where the intervention should impact both level and rate of improvement, B trend should be included in the effect size.

As an exposition and field test of Tau-U, this article has several limitations. First, this is a substantially new model, and parallels drawn to regression may seem stretched. For example, for a given AB dataset, results (SS_{model}, R and R^2) from a simple mean shift (SMS) regression model, will always be smaller than those from a mean + trend model (MTS). That is not the case with Tau-U; adding trend can easily drop Tau-U values. That is because in regression both models reference the same SS_{tot}, whereas in Tau-U's S variance model, the total number of pairs (analogous to total variance) varies depending on which partitions of the matrix are included. This fundamental difference between an ANOVA variance matrix and the S difference matrix may take some getting used to.

Another limitation was the field test to demonstrate Tau-U's baseline control effects. It included only one set of analyses on a sample of 176, and lacked a graphic display to demonstrate the impact of baseline trend control. To date we have not been able to construct such a display. A related limitation is that Tau-U trend control was compared only with the Allison et al. semipartialling approach (Allison & Gorman, 1993; Faith et al., 1997). Other variance-based trend control techniques are now being developed for growth modeling within multilevel models (MLM) and structural equation models (SEM). They were not included in this paper, as they have not yet been adequately proved with real SCR data, as the Allison model has.

Although linear regression is still an important method for SCR analysis, maximum likelihood

Please cite this article as: Richard I. Parker, et al., Combining Nonoverlap and Trend for Single-Case Research: Tau-U, *Behavior Therapy* (2011), 10.1016/j.beth.2010.08.006

algorithms may side step most ordinary least squares data assumptions. Furthermore, there is much recent development in nonparametric trend analysis, which can handle data with only ordinal level properties. The field is moving quickly. Given the likelihood that multilevel modeling and possibly structural equation modeling will be successfully adapted to single-case research in the near future, our primary concern is that the effects of those analyses on an individual client's data graph be made transparent. Validation by visual analysis is especially important with increasingly complex analyses.

A final limitation is that Tau-U's application to more complex designs (which predominate the literature) was not demonstrated. We do not anticipate difficulty in doing so; the most attractive and generally usable technique seems to be through meta-analysis software, in which each AB contrast is a separate strata within a fixed-effects model. Free downloadable software such as WinPEPI (Abramson, 2010) automatically weights results for each series by the inverse of its variance, to obtain an omnibus effect size with narrower confidence intervals.

In summary, Tau-U is an index with more statistical power than any other nonoverlap (dominance) index known. It also is the most discriminating, by not hitting the 100% nonoverlap ceiling that challenges much SCR research. The distribution of Tau-U is nearly ideal, like regression analyses, and unlike simple nonoverlap. Tau-U is flexible in that it can calculate trend only, nonoverlap between phases only, or a combination. Its abilities to include Phase B trend and to control unwanted Phase A trend parallel the flexibility of regression. However, the Tau-U control method is unique, and may seem strange to those familiar with regression. The net effect of controlling Phase A trend is conservative, causing a smaller impact on results than we are used to seeing in regression. And the net effect of adding B trend is an estimate of trend within and across phases that tends to be smaller than simple nonoverlap.

Tau-U can be calculated from any KRC module that provides Kendall's *S*, also known as "*S*" or "score," along with *p* values. Unfortunately, SPSS does not, but SAS does. For small data sets, exact permutation-based *p* values are also desirable. Remember that the KRC module was not built for dummy-coded data, so the Tau, #pairs, and #ties output will not be accurate. *S* will be accurate, however, as will its *p* values, standard error of *S*, and variance of *S*. The user must hand calculate #pairs, since *S* / #pairs = Tau. Recall that for an AB phase contrast, #pairs = $n_A \times n_B$. If B trend is added, the additional #pairs = $n_B \times n_B - 1 / 2$. The most convenient analytic tool may be the free Web-based KRC module by Wessa (2008) at http://www.wessa.net/rwasp_kendall.wasp/. The Web-based Wessa software (developed from open-source "*R*") outputs accurate *S*, *S* variance, and exact *p* values. The software with the most complete output we have found is StatsDirect Ltd. (2010), inexpensive software from Great Britain for medical researchers, with extensive nonparametric capabilities. Most analyses in this paper were by StatsDirect.

As this article goes to press, we have just completed a "stand alone" statistical application for calculating Tau-U on more complex designs. It is web-based and will be made freely available. Readers can contact the second author for its web site, which will be available within weeks from now.

References

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

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.

Armitage, P., Berry, G., & Matthews, J. N. (2002). *Statistical Methods in Medical Research*, 4th ed. Oxford: Blackwell Science.

Borckardt, J. J., Nash, M. R., Murphy, M. D., Moore, M., Shaw, D., & O-Neil, P. (2008). Clinical practice as natural laboratory for psychotherapy research. *American Psychologist*, *63*, 77–95.

Box, G. E. P., & Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control (Rev. ed.)*. Holden-Day: San Francisco.

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

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

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

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.

Cochrane, D., & Orcutt, G. H. (1949). Application of least squares regression to relationships containing autocorrelated error terms. *Journal of the American Statistical Association*, *44*, 32–61.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*, 2nd ed. Erlbaum: Hillsdale, NJ.

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

Cooper, J. O., Heron, T. E., & Heward, W. L. (1987). *Applied behavior analysis*. Columbus, OH: Merrill.

Cottrell, A., & Lucchetti, R. (2009). *GNU regression, econometric, and time-series library* [Computer software]. Retrieved from http://gretl.sourceforge.net/.

Please cite this article as: Richard I. Parker, et al., Combining Nonoverlap and Trend for Single-Case Research: Tau-U, *Behavior Therapy* (2011), 10.1016/j.beth.2010.08.006

Crosbie, J. (1993). Interrupted time-series analysis with brief single-subject data. *Journal of Consulting and Clinical Psychology*, *61*(6), 966–974.

Crosbie, J. (1995). Interrupted time-series analysis with short series: Why it is problematic; How it can be improved. In J. M. Gottman (Ed.), *The analysis of change*. Mahwah, NJ: Erlbaum.

Faith, M. S., Allison, D. B., & Gorman, B. S. (1997). 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: Erlbaum.

Franklin, R. D., Allison, D. B., & Gorman, B. S. (1997). Introduction. In D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 1–12). Mahwah, NJ: Erlbaum.

Glass, G. V., Willson, V. L., & Gottman, J. M. (1975). *Design and analysis of time series experiments*. Boulder: University of Colorado Press.

Harvey, A. C. (1981). *The econometric analysis of time series*. New York: Wiley.

Harvey, A. C., & McAvinchey, I. D. (1978). *The small sample efficiency of two-step estimators in regression models with autoregressive disturbances (Paper No. 78-10). Vancouver*. Canada: University of British Columbia.

Hildreth, C., & Lu, J. Y. (1960). *Demand relations with autocorrelated disturbances (Technical Bulletin No. 276)*. East Lansing: Michigan State University.

Hintze, J. (2006). *NCSS and PASS : Number cruncher statistical systems* [Computer software]. Kaysville, UT.

Hollander, M., & Wolfe, D. A. (1999). *Non-parametric statistical methods*, 2nd ed. New York: Wiley.

Huberty, C. J., & Lowman, L. L. (2000). Group overlap as a basis for effect size. *Educational and Psychological Measurement*, *60*(4), 543–563.

Huitema, B. (2004). Analysis of interrupted time-series experiments using ITSE: A critique. *Understanding Statistics*, *3*, 27–46.

Johnston, J. M., & Pennypacker, H. S. (1993). *Strategies and tactics of behavioral research*. Hillsdale, NJ: Erlbaum.

Jones, R. R., Vaught, R. S., & Weinrott, M. (1977). Time-series analysis in operant research. *Journal of Applied Behavior Analysis*, *10*, 151–166.

Judge, G. G., Griffiths, W. E., Hill, R. C., & Lee, T. C. (1985). *The theory and practice of econometrics*, 2nd ed. New York: Wiley.

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

Kendall, M. G., & Gibbons, J. D. (1999). *Rank correlation methods*, 5th ed. London: Arnold.

Kratochwill, T. R., & Levin, J. R (Eds.). (1992). *Single-case research design and analysis: New directions for psychology and education*. Hillsdale, NJ: Erlbaum.

Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). *Applied linear regression models*, 4th ed. New York: McGraw-Hill.

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: Erlbaum.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models*, 4th ed. Boston: McGraw-Hill.

Park, R. E., & Mitchell, B. M. (1980). Estimating the autocorrelated error model with trended data. *Journal of Econometrics*, *13*, 185–201.

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

Parker, R. I., & Brossart, D. F. (2006). Phase contrasts for multi-phase single case intervention designs. *School Psychology Quarterly*, *21*(1), 46–61.

Parker, R. I., Brossart, D. F., Callicott, K. J., Long, J. R., De-Alba, R. G., 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*(3), 418–440.

Parker, R. I., & Vannest, K. J. (2009). An improved effect size for single case research: Non-overlap of all pairs (NAP). *Behavior Therapy*, *40*(4), 357–367.

Parker, R. I., Vannest, K. J., & Brown, L. (2009). The improvement rate difference for single case research. *Exceptional Children*, *75*(2), 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: Erlbaum.

Prais, S. J., & Winsten, C. B. (1954). *Trend estimators and serial correlation (Report No. 383)*. Chicago: Cowles Commission.

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. (2001). How to summarize single-participant research: Ideas and applications. *Exceptionality*, *9*, 227–244.

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

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

Southerly, B. (2006, February 14). *RE: ITSACORR update* [Online forum comment]. Retrieved from http://www.mail-archive.com/tips@acsun.frostburg.edu/msg16062.html.

Spitzer, J. J. (1979). Small-sample properties of nonlinear least squares and maximum likelihood estimators in the context of autocorrelated errors. *Journal of the American Statistical Association*, *74*(365), 41–47.

Sprent, P., & Smeeton, N. C. (2007). *Applied nonparametric statistical methods*, 4th ed. New York: Chapman & Hall/CRC.

StatsDirect Ltd. (2010). *StatsDirect statistical software* [Computer software]. London. Retrieved from http://www.statsdirect.com.

Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data*. Hillsdale, NJ: Erlbaum.

Wessa, P. (2008). *Kendall Tau Rank Correlation—free statistics software, Version 1.1.23–r6* [Computer software]. Office for Research Development and Education. Retrieved from http://www.wessa.net/rwasp_kendall.wasp/.

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

Wilcox, R. (2001). *Fundamentals of modern statistical analysis: Substantially improving power and accuracy*. New York: Springer-Verlag.

RECEIVED: December 19, 2009
ACCEPTED: August 4, 2010

Please cite this article as: Richard I. Parker, et al., Combining Nonoverlap and Trend for Single-Case Research: Tau-U, *Behavior Therapy* (2011), 10.1016/j.beth.2010.08.006