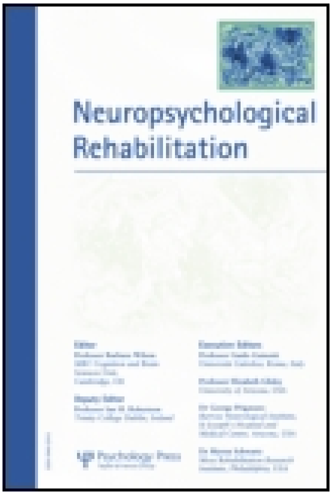


Neuropsychological Rehabilitation: An International Journal
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/pnrh20

Click for updates

Incorporating nonoverlap indices with visual analysis for quantifying intervention effectiveness in single-case experimental designs

Daniel F. Brossart[a], Kimberly J. Vannest[a], John L. Davis[a] & Marc A. Patience[a]

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

Published online: 05 Feb 2014.

To cite this article: Daniel F. Brossart, Kimberly J. Vannest, John L. Davis & Marc A. Patience (2014) Incorporating nonoverlap indices with visual analysis for quantifying intervention effectiveness in single-case experimental designs, Neuropsychological Rehabilitation: An International Journal, 24:3-4, 464-491, DOI: 10.1080/09602011.2013.868361

To link to this article: http://dx.doi.org/10.1080/09602011.2013.868361

Neuropsychological Rehabilitation, 2014
Vol. 24, Nos. 3–4, 464–491, http://dx.doi.org/10.1080/09602011.2013.868361

Routledge
Taylor & Francis Group

Incorporating nonoverlap indices with visual analysis for quantifying intervention effectiveness in single-case experimental designs

Daniel F. Brossart, Kimberly J. Vannest, John L. Davis, and Marc A. Patience
Texas A&M University, College Station, TX, USA

(*Received June 2013; accepted November 2013*)

The field of neuropsychological rehabilitation frequently employs single case experimental designs (SCED) in research, but few if any, of the published studies use the effect sizes recommended by the American Psychological Association. Among the available methods for analysing single case designs, this paper focuses on nonoverlap methods. This paper provides examples and suggestions for integrating visual and statistical analysis, pointing out where contradictions may occur and how to be a critical consumer.

Keywords: Nonoverlap; Visual analysis; Effect size; Single-case; Tau-U.

INTRODUCTION

The advantages and disadvantages of single case experimental designs (SCEDs) have been noted in numerous books, chapters, and articles across multiple disciplines. This discussion is often framed in such a way that SCEDs are compared to randomised clinical trials (RCTs). This effort, one may presume, is to show that SCEDs have important strengths that RCTs do not. Suffice it to say that single-case designs can be among “the most

Correspondence should be addressed to Daniel F. Brossart, 4225 TAMU, College Station, TX 77843-4225. E-mail: brossart@tamu.edu

The authors would like to thank Kevin R. Tarlow for creating the R code used to calculate all variations of Tau-U.

effective and powerful" (Shadish, Cook, & Campbell, 2002, p. 171) nonrandomised experimental designs (Shadish, Rindskopf, & Hedges, 2008).

SCEDS, CLINICAL PRACTICE, AND CONCEPTUAL FOUNDATIONS

Arguably, SCEDs can be integrated into many clinical practices where one is concerned about a patient's response to treatment (Barnett et al., 2012). Instead of using RCTs to search for effective treatments, some have championed the benefits of beginning with research that identifies effective practice in the community – studying those interventions used by experienced clinicians (Blais & Hilsenroth, 2007). SCEDs could play an important role in such an endeavour. Relatedly, some contend that clinicians should be a major force in developing evidence for practice (Grimmer, Bialocerkowski, Kumar, & Milanese, 2004). This would help ensure that research is meaningful for the clinician, eventually developing into partnerships with investigators and resulting in clinically meaningful research that would improve the treatments delivered by clinicians.

The need for evidence-based practice has been articulated for numerous disciplines in statements produced by organisations such as the Cochrane Collaboration, What Works Clearinghouse, Campbell Collaboration, Coalition for Evidence-Based Policy, and the American Speech-Language-Hearing Association's National Center for Evidence-Based Practice in Communication Disorders. In addition, many organisations are involved in producing guidelines for practice that are designed to assist the clinician or practitioner in making treatment decisions. Evidence-based practice places the focus on the individual patient. Yet for research findings to be adopted and to impact practice, such findings must have clinical relevance. Clinicians are usually concerned about the treatment of their patient, whereas many research endeavours provide information on how a treatment impacts a sample. In many cases, there is no clear linkage between the treatment of their particular patient and how a sample responded to a treatment.

Part of the problem in linking scientific findings to the treatment of a particular patient may be due to the overuse of one kind of research paradigm. Some have framed the issue as one of needing to adopt a more idiographic approach in lieu of the nomothetic approach (e.g., Lamiell, 1981). Yet the issue is more complicated than what on the surface seems to be simply studying individuals rather than groups. Idiographic knowledge is concerned with describing and explaining particular phenomena, while nomothetic knowledge is concerned with "finding generalities that are common to a class of particulars and deriving theories or laws to account for these generalities" (Robinson, 2011, p. 32). Historically it has been argued that these are complementary sides of all science rather than opposing perspectives. Yet as noted by Robinson (2011) there are two forms of the nomothetic approach. One seeks to determine what is common to all individuals in a sample or category (the Wundtian model) and the other seeks to determine what is common to a sample or group of persons as an aggregated whole (the Galtonian model). The Wundtian approach would use individual cases to develop and test theory (often case-by-case). The goal would be to determine what is "common to all", where any case not conforming would be a challenge to the theory. Such a process is markedly different than making statements about what is true for an aggregated whole.

The term idiographic has been used recently to refer to a variety of things beyond its original definition. For example, Barlow and Nock (2009) seem to associate equivalence with SCEDs and idiographic research. Others, such as Molenaar (2004) and Nesselroade (1991) focus on issues of variability (intraindividual and interindividual variability), while Cattell (1988) used a Data Box or Basic Data Relations Matrix to illustrate how differing mixtures of variables, persons, and occasions provide different ways to address various research questions ranging from those complementary to the Galtonian model (such as R technique, which consists of many variables in columns, many people in rows, one occasion) to those more in line with a Wundtian model (e.g., P technique, which consists of many variables, many occasions, one person). As noted by Robinson (2011), idiographic research is not tied to a specific method, it is an objective – the objective to describe or explain a single phenomenon.

While some would see adopting a Wundtian approach or using SCEDs as a step in the right direction in terms of balancing the overuse of the Galtonian approach, it does not automatically solve many issues facing the researcher seeking to produce clinically relevant results. For instance, the proper way to analyse data from SCEDs remains a topic under investigation, with no single approach proving to be superior in every instance. Historically, visual analysis was viewed as the proper way to analyse such data. Multiple studies have examined the reliability of human judges analysing single-case data and have generally found that human judges are mediocre at best even when given contextual data in which to interpret the graphs (e.g., Brossart, Parker, Olson, & Mahadevan, 2006). It appears that the one thing that human judges are reasonably good at is detecting graphs that show no treatment effect (Ximenes, Manolov, Solanas, & Quera, 2009). A few studies have reported high inter-rater agreement, but these studies have important interpretive aids. For instance, one study involved multiple phases (ABAB), and raters were asked to make a determination if the graph demonstrated experimental control on a scale of 0 to 100 (Kahng et al., 2010). Another study (Hagopian, Fisher, Thompson, & Owen-DeSchryver, 1997) had the participants draw upper and lower criterion lines (approximately 1 *SD* from the mean of the control condition). Rules were then given on how to interpret the graphs based on these lines. A high level of agreement was achieved after the participants underwent training on using the rules for the interpretation of multi-element functional analysis data. Research is ongoing in regards to the effects various graphing characteristics have on visual judgement. In some instances, the results suggest that some graphing conventions may have an impact on the Type I and Type II error rates, but more research is needed (Carter, 2009).

STATISTICAL ANALYSIS, EFFECT SIZES, AND VISUAL ANALYSIS

Most researchers investigating statistical methods to analyse single-case data advocate the use of both visual and statistical analysis such that both inform the other (Brossart et al., 2006; Franklin, Allison, & Gorman, 1996). One difficulty the investigator faces is choosing from the bewildering number of statistical methods available (Parker & Brossart, 2003). There are a number of non-parametric effect sizes that produce an effect size ranging from 0 to 1 that provide an estimate of the size of change in response to a treatment or intervention.

Use of non-parametric analysis is important because single case studies typically have short data sets or few data points, non-normal or unknown distributions, and unknown parameters. When a confidence interval and *p* value are provided with an effect size, the results give us an estimate of the probability of chance occurrence of the effect size and an explicit estimate of the error based on a confidence level we find desirable. There are many non-parametric effect sizes in the published literature, each with their relative strengths and weaknesses.

An early effect size was the extended celeration line (ECL; White & Haring, 1980) followed by percentage of non-overlapping data (PND; Scruggs, Mastropieri, & Casto, 1987), percent of data exceeding the median (PEM; Ma, 2006), percentage of all non-overlapping pairs (PAND; Parker, Hagan-Burke, & Vannest, 2007), non-overlap of all pairs (NAP; Parker & Vannest, 2009), the improvement rate difference (IRD; Parker, Vannest, & Brown, 2009), the percent of data exceeding the phase A median trend (PEM-T; Wolery, Busick, Reichow, & Barton, 2010), and, finally, Tau-U (Parker, Vannest, Davis, & Sauber, 2011b).

Existing papers present the techniques and one paper reviews them in more depth (Parker, Vannest, & Davis, 2011a) so we will not attempt to do that here, but rather to demonstrate how each performs by first providing a brief review of each technique using an artificial data set to demonstrate differences in procedures, results, strengths and weaknesses. Then we use existing data to evaluate how the superior indices perform with typical neuropsychological data. We conclude with some examples of when contradictions may occur so that the reader may make informed choices and thoughtful interpretations.

Effect size indices should ideally be used in conjunction with parts of a visual analysis, perhaps better termed “design analysis”. A design analysis is the notion of assessing the SCED for threats to internal or conclusion validity and we introduce the term here to distinguish a design analysis from a visual analysis. A design analysis examines the structure of the experiment to see if it is valid for asserting a functional relationship, whereas a visual analysis looks at means, trends, immediacy and consistency of change in assessing type and amount of behaviour change. This visual analysis may inform a determination about a functional relationship, but not in isolation from evaluation of the design. For example, an AB design alone can never be used to determine causation.

A visual analysis looks for effects of treatment by examining relative size of change, onset of change, trend or stability of measurement, replication demonstrations and consistency. A design analysis looks only at the adequacy of a design for conclusion validity. It is the design of the study that controls for threats to validity (Kazdin, 2011; Kennedy, 2005). A study with poor internal validity or poor conclusion validity will not allow one to make strong inferences about its findings. Therefore a design analysis occurs before effect size calculation by determining if a functional relationship was established. The criterion for a functional relationship includes three or more demonstrations of experimental control or phase changes (Kratochwill et al., 2010). For example an ABAB design would show behaviour change between A1B1, B1A2, and A2B2. Another example would be a multiple baseline design with behaviour change across three or more participants or settings. A design analysis can be applied a priori to a study or post-hoc in the case of a meta-analysis.

When calculating effect sizes for a meta-analysis or to determine evidence-based practices, another concept is required and that is to examine the quality of the studies included (Horner et al., 2005; Kratochwill et al., 2010). Things like careful consideration of participant description and sufficient procedural detail for replication improves the external validity of the study. Quantification of SCED should only occur when prior conditions for quality are met (Horner et al., 2005; Kratochwill et al., 2010).

REVIEW OF EFFECT SIZE INDICES

We are frequently asked, “Which effect size should we use?” Although some may assume it is just a matter of selecting the largest one, prior research suggests the matter is more complicated than basing one’s decision on the size of the effect. So how does one determine which effect size is a best performer? Several methods have been documented to have important characteristics that should be considered before their use. PND has well-documented limitations and several researchers recommend it not be used (Kratochwill et al., 2010; Parker & Vannest, 2009). PEM has performed poorly when compared to other methods; it was unable to discriminate among data sets that posed no problem for other methods (Parker, Vannest, & Davis, 2011a; see also Wolery et al., 2010). Other studies have found PAND to be problematic in that it yielded similar results for data sets with and without treatment effects (Manolov, Solanas, & Leiva, 2010). In terms of relative power, ECL and PEM appear to have the lowest power among non-overlap methods. IRD has moderate power, but variability causes a problem in sensitivity (Manolov, Solanas, Sierra, & Evans, 2011). IRD and PNCD (a newer and less studied variation of PND, Manolov & Solanas, 2009) appear to be somewhat dependent on the length of the series analysed (Manolov & Solanas, 2009; Manolov et al., 2011). Those with the most power are NAP, and Tau-U (Parker, Vannest, & Davis, 2011a). NAP appears to perform adequately under higher levels of autocorrelation and is unaffected by distortions due to heteroscedasticity except when there was an exponential uniform random variable term in the generated data in a simulation study (Manolov et al., 2011). Yet the same group of investigators found that linear trend inflated NAP and to a greater extent IRD. Even so, it appears that NAP is one of the better performing indices. Tau-U is an extension of NAP with the ability to correct for trend (PNCD also controls for baseline trend, Manolov et al., 2011). Tau-U addresses change in trend and level, it is distribution free, is only somewhat influenced by autocorrelation, and when controlling for trend, does so in a more conceptually defensible manner than the regression method advocated by Allison and colleagues (Parker, Vannest, Davis, & Sauber, 2011b).

In addition to general strengths and limitations of each statistical method, the single-case investigator will often be faced with data possessing certain characteristics that make analysis more difficult. Small data sets, variability, inconsistent effects, and small or gradual behaviour change are just some of the problems encountered in SCEDs. Another is the presence of baseline trend. Correcting for baseline trend is a practice that continues to be debated. If trend is present and arguably needs to be controlled, then only three non-overlap methods are currently able to do so. PNCD controls for baseline trend by a differencing procedure before the intervention effect is calculated (Manolov & Solanas, 2009), but if no trend is present in the baseline phase then other methods may be preferred. ECL controls for trend that is assumed to be linear and to continue into the treatment phase. Tau-U controls for monotonic trend. Thus, for many situations it appears that Tau-U is the better performing non-parametric method for analysing single-case data currently available. A summary of six non-overlap methods is presented in Table 1, which contains a brief overview of how the statistic is calculated, and strengths and weaknesses of each method are included.

Nevertheless, the best approach may be to report multiple effect sizes similar to the practice of reporting multiple fit indices in structural equation modelling (SEM). In SEM, multiple fit indices are typically reported because each fit index conveys something different than the other fit indices. Some fit indices are absolute measures of fit whereas others are incremental or comparative fit indices. Some contain a penalty based on the number of parameters estimated, while others do not. In similar fashion, for reporting single-case effect sizes, the best approach may be to report several effect sizes since each method appears to have strengths and weaknesses.

In the remainder of the paper we report both IRD and Tau-U. We felt it important for two methods to be used in our examples to provide some means of comparison. IRD, though limited in some aspects, is easy to compute and seemed a reasonable method to compare with Tau-U. IRD is from the risk ratio family of effect sizes and is theoretically different (a proportion) than Tau-U which is a dominance statistic.

HAND CALCULATION OF TAU

Here we present an artificial data set to compare results and review strengths and limitations of each technique. Given a sample data set for an ABAB reversal design where phase A is baseline and phase B is treatment (Figure 1). Consider the following data for "Jack" A1 – 3, 2.5, 3.5, 7, 3, 3; B1 – 8, 7, 6.5, 8, 8, 6, 5, 9, 10; A2 – 7, 6, 5, 3, 3, 5, 2, 3, 4, 2; B2 – 8, 9, 8, 10, 10, 9, 10, 8, 10, 9.5, 9, 8, 8, 8.4, 7.8. The example data analysis will be for the first AB phase contrast only.

Although complete steps are available in prior publications (e.g., Parker et al., 2011b), the heuristic below demonstrates the hand procedure in greater depth than that provided in Table 1 for calculating Tau-U

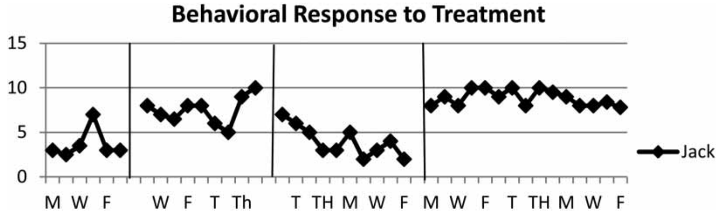
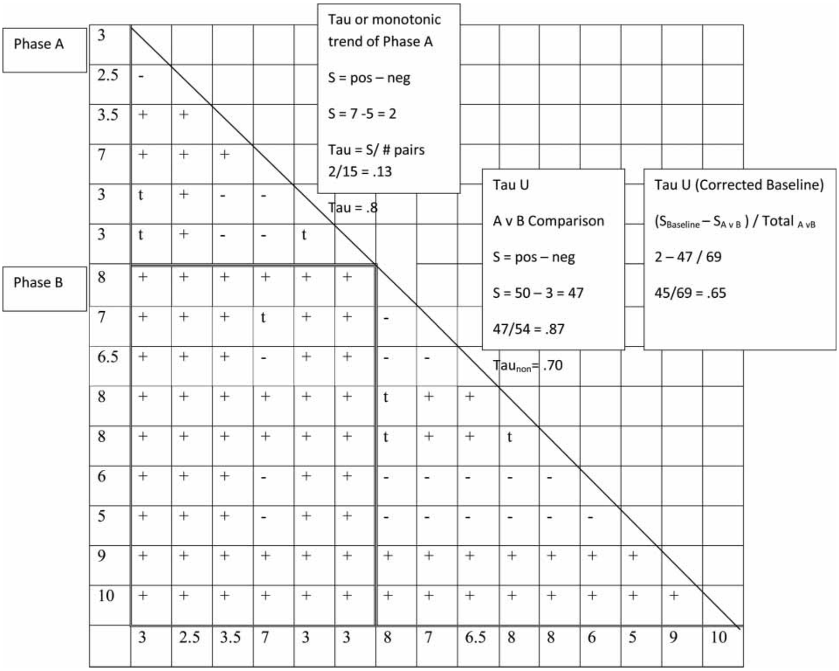


Figure 1. Graph of data set for Jack.

TABLE 1
Comparison of effect size calculation steps for methods by hand using same data set

	PND	ECL	PEM	IRD	NAP	TauU/Tau
Step 1	Identify the number of data points in intervention exceeding the highest point in baseline (5)	Calculate the median trend line of phase A and extend through phase B	Find the median of phase A and extend the line through phase B	Identify and remove the fewest number of data points to eliminate the overlap (1)	Compare each pairwise comparison and assign value of 1 for improvement, 0.5 for tie or 0 for no improvement	Compare each data point in forward manner using a matrix, score +, t, −
Step 2	Divide the higher number by the total number of data points in phase B 5/9 = .55	Divide the number of data points exceeding the line by the total number of data points 10/10 = 1	Divide the number of data points in B exceeding the line by the total number of data points 8/10 = .80	Data improved in phase B (9/9 = 1). Data improved in phase A (1/6 = .17). Subtract 1−.17 = .83	Number of all possible pairs in phase A data points x phase B data points (6×9 = 54). Number of improvement divided by total is 50.5/54 = .93	Calculate S/# pairs where S = pos-neg and # pairs is number of pair-wise comparisons 47/54 = .87
Step 3	Result is .55 or 55%	Recalibrate score (1×2) − 1 = 1 or 100%	Recalibrate score (.80 x 2) − 1 = .60 or 60%	Result is .83 or 83%	Recalibrate (.93 x 2) − 1 = .86 or 86%	.87%
	Strengths					
	-Field tested in dozens of studies for more than a decade. -Easy hand calculation	-Long history. -Adjusted for trend. -Easy hand calculation.	-More representative than PND by using Mean	-Well used in medical literature.	-Superior precision and power. -Direct calculation and interpretation	-Ability to control trend. -All data involved. -Conservative but not overly so.
	Limitations					
	-Floor and ceiling effects. -Lacks sampling distribution. -Will not work for a meta-analysis.	-Lack of precision and power. -Assumes linearity. -Unreliable phase a trend.	-Low power. -Lack of sensitivity. -Severe ceiling effects.	-Insensitive to trend.	-Insensitive to trend. -Not as easy to calculate by hand.	-May not show sensitivity to all data patterns. -Not easy by hand on long data sets.



Phase A	3															
	2.5	-														
	3.5	+	+													
	7	+	+	+												
	3	t	+	-	-											
	3	t	+	-	-	t										
Phase B	8	+	+	+	+	+	+									
	7	+	+	+	t	+	+	-								
	6.5	+	+	+	-	+	+	-	-							
	8	+	+	+	+	+	+	t	+	+						
	8	+	+	+	+	+	+	t	+	+	t					
	6	+	+	+	-	+	+	-	-	-	-	-				
	5	+	+	+	-	+	+	-	-	-	-	-	-			
	9	+	+	+	+	+	+	+	+	+	+	+	+	+		
	10	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
		3	2.5	3.5	7	3	3	8	7	6.5	8	8	6	5	9	10

Tau or monotonic trend of Phase A

S = pos – neg

S = 7 -5 = 2

Tau = S/ # pairs
2/15 = .13

Tau = .8

Tau U

A v B Comparison

S = pos – neg

S = 50 – 3 = 47

47/54 = .87

Tau_{non}= .70

Tau U (Corrected Baseline)

$(S_{\text{Baseline}} - S_{A\,v\,B}) / \text{Total}_{A\,vB}$

2 – 47 / 69

45/69 = .65

Figure 2. Example matrix for calculating Tau by hand.

(Figure 2). The steps for calculating Tau-U are as follows: each data point compared to all data points ahead of it in time in a “forward” direction. For each pairwise comparison determine if the earlier data point is larger (assign a +) smaller, (assign a –), or equal (assign a t for tie). To determine the number of possible comparisons multiply the number of data points in phase A with the number in phase B (N* (N–1))/2. To calculate this by hand and summarise these data visually, create a half-matrix where phase A and phase B data are enumerated on the X and Y axis and the “+”, “–”, or “t” is assigned to each cell.

The data in this matrix should form a triangle with the data forming three regions. The triangles represent the trend within the A or B phase and the rectangle is the A vs. B nonoverlap. To determine the Tau in phase A add the values in that triangle and divide by the comparisons in that part of the matrix. The process is the same to determine the Tau in phase B. Tau-U can be applied as the primary analysis of the AB comparison of nonoverlap, but also within phases to determine trend when combined trend can be conservatively removed from either side A or B or both. More detail on controlling for baseline trend and accounting for phase B trend is reported in Parker et al. (2011b). These variations of Tau-U include: (a) AB phase comparison, (b) AB comparison controlling for phase A trend, (c) AB comparison plus phase B trend, (d) AB comparison controlling for phase A trend plus phase B trend. These variations make Tau-U very versatile, but also require conceptual clarity on the user's part to report the output that makes the most sense given their particular application and data characteristics.

To calculate Tau-U one may prefer to use R syntax developed to calculate Tau-U (https://dl.dropboxusercontent.com/u/2842869/Tau_U.R). To use one's own data, create a comma delimited file with time in the first column, one's data in the second column, and phase in the third column using 0 or 1. The first row should contain the variable name for each column. The R syntax requires that the Kendall package be loaded in R. This syntax will produce output for each variation of Tau-U described in Parker et al. (2011b).

ILLUSTRATIVE ARTICLES

In order to obtain single-case graphs for the purpose of evaluating effect size calculation methods, a review of articles from the journal *Neuropsychological Rehabilitation* published between 1987 and 2013 was performed. A total of 123 single case graphs were pulled from 32 articles. Graphs were saved from pdf files using the Snipping tool. After a review of articles, two independent raters identified data sets which represented a range of common configurations in data (e.g., positive baseline trend, short data phases). We also selected more complicated or unusual data configurations which may pose challenges to analysis and interpretation. This created a sample of real (vs. Monte Carlo simulation) data for comparison across methods. Two additional raters were then asked to independently agree or disagree for consensus with the representative descriptions of the graphs as to whether they met inclusion criteria to "demonstrate strengths and weaknesses of the various analyses", simple agreement was 100%. The final sample included 15 graphs representing common variations in data sets which might affect effect size selection and results. Raw data are typically represented graphically in SCR studies. Given these original data, it is possible to convert these graphical data to exact numerical form using digital conversion methods. Software packages are widely available to perform this task. The availability of raw data in SCR studies provides an opportunity to calculate any post hoc effect size.

Data extraction

For the current study, graphic data from published studies were extracted and digitised in several steps. Details of the data extraction process are as follows. First, graphs from published studies were saved as pdf files. Next, each graph was copied from the published articles using the Snipping tool from Microsoft Windows 7. This digital snapshot tool captures just the image of the graphic data which can then be uploaded into the digitising program (GetData Graph Digitizer, Version 2.25, http://getdata-graph-digitizer.com). Next, each image is loaded individually, and the scale for the X and Y axis must be set to match the axis values within the graph. Following this procedure, each data point is specified to assure exact concordance with the original study data.

Using these data, effect sizes were calculated by hand for IRD and using the R syntax available online. Both effect size indices were calculated for each graph. Confidence intervals and *p* values were calculated for Tau-U. If multiple cases were present on an image, one case was chosen.

COMPARISON OF STATISTICAL EFFECT SIZE TO VISUAL INTERPRETATION

Because we hoped to compare visual judgements about magnitude of effect with a statistical effect size, we then had each of the 14 graphs visually analysed, first by one class of 13 doctoral (PhD) students and next by another class of eight masters (MA) students. All graduate students had experience in single case research methodology and design, including two or more courses, clinic or field work, and were at the end of the semester in an advance course on SCED. Each graph was presented as a Power Point slide and students were presented one slide at a time and asked to characterise the effect size as small, medium, or large. Students were specifically instructed to analyse the AB contrast holistically rather than by attempting calculations in their heads or by use of a single dimension such as variability or overlap of data. The evaluation and recording for each student was done independently. After visual analysis, results were compounded and are presented below each graph, alongside the effect size calculations.

RESULTS

A review of 25 years of *Neuropsychological Rehabilitation* produced 32 SCED studies and 15 demonstrations of “types” of data. IRD and Tau-U effect sizes were run for each data set using an AB contrast. IRD *p* values may be produced in many statistical packages by the feature that calculates the difference between two proportions, under “proportion statistics” or “risk analysis” (Parker et al., 2009). Tau-U is derived from the Kendall Rank Correlation and the Mann-Whitney *U*, both of which rely on the *S* distribution for significance testing (see Parker et al., 2011b for more information on calculating significance values).

Where multiple data series are presented, the first data series was used in the calculation of effect size and only the first baseline and intervention phase were analysed if multiple phases were present. Effect sizes, *p* values, and the visual analysis interpretations from masters and doctoral students trained in SCED are reported as a frequency count of those judges who rated the effect size as a small, medium, or large effect. For brevity we selected 10 illustrative graphs.

The first three figures (Figures 3, 4, and 5) demonstrate data which are variable to some extent (Figure 3), sparse and overlapping (Figure 4), or have a clear trend (Figure 5). Yet in each of these three cases, there is relatively high agreement between two types of statistical analysis and independent visual analysis. For example, IRD and Tau-U in Figure 3 were .15 and .09 (both small) with no statistical significance for the effect. IRD and Tau-U for Figure 4 were .30 and .43 with no statistical significance; IRD and Tau-U for Figure 5 were .86 and .91 (large) with statistical significance. Likewise, the majority of the visual analyses agreed with the statistical analysis. Figures 3 and 4 were both rated by the majority as demonstrating small effects and Figure 5 was rated as large. These examples serve to illustrate that statistical analysis can be congruent with the decision making of most visual analysts and may serve to clarify interpretation of effects when visual analysts disagree by providing additional information such as statistical significance.

Figure 6 represents another common data scenario, few data points. Both IRD and Tau-U produced an effect size of 1, but the *p* value tells us the result

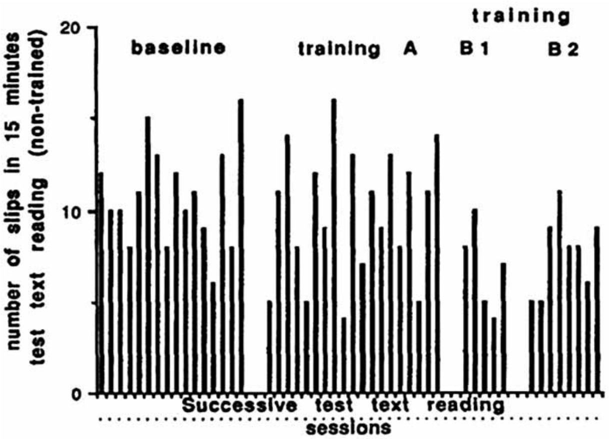


Figure 3. From Wilson, C., & Roberston, I. H. (1992). A home-based intervention for attentional slips during reading following head injury: A single case study. *Neuropsychological Rehabilitation, 2*(3), 193–205. Reprinted with permission. IRD = .15; Tau-U = .09; *p* =.68; Visual analysis ES (S, M, L) MA = 4, 1, 3; PhD = 8, 2, 0.

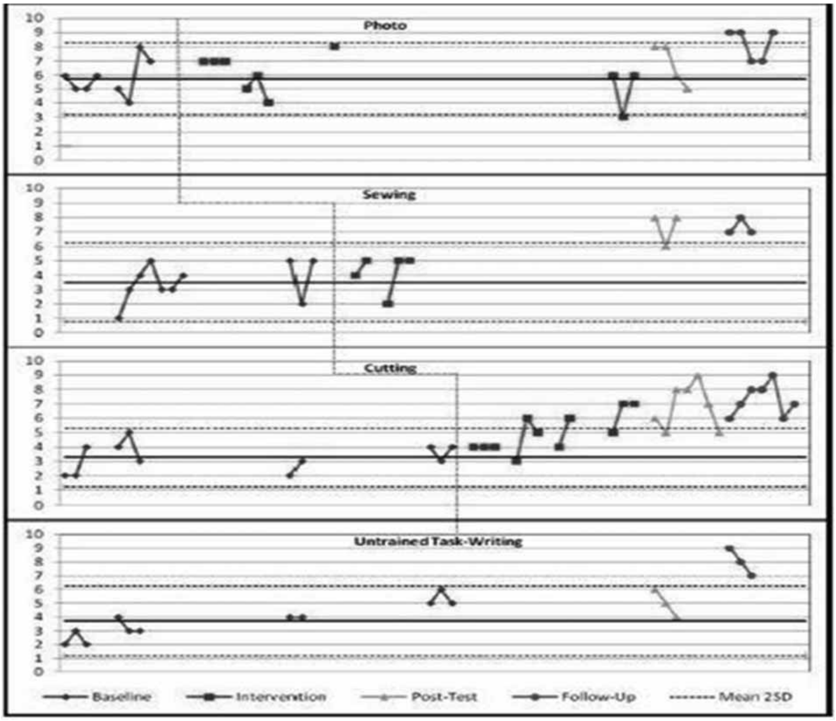


Figure 4. From McEwen, S. E., Polatajko, H. J., Huijbregts, M. P. J., & Ryan, J. D. (2010). Inter-task transfer of meaningful, functional skills following a cognitive-based treatment: Results of three multiple baseline design experiments in adults with chronic stroke. *Neuropsychological Rehabilitation, 20*(4), 541–561. Reprinted with permission. IRD = .30; Tau-U = .43; p =.08; Visual Analysis ES (S, M, L) MA = 8, 0, 0; PhD = 9, 3, 0.

is not significant. This is a case where a p value for the effect size can inform visual analysis. Eleven of 18 graduate students rated the effect as large, five as medium, and two as small. When the two who ranked small were queried, they said, “too few data points to be certain about the effects”, essentially moderating the effect size with a human-determined significance test of chance.

Using the top series in Figure 7, it is presented as it was in the journal with a trend line. Visible trend lines can skew graph interpretation and decision making (DeProspero & Cohen, 1979; Greenspan & Fisch, 1992; Hojem & Ottenbacker, 1988; Skiba, Deno, Marston, & Casey, 1989). IRD and Tau-U effect sizes were .12 and .43, p = .07, demonstrating very different results. Eleven of 18 visual analysts (61%) found the effect to be medium, three rated the effect as small and four as large. Figure 8 also shows data represented by a trend line. Treatment demonstrates a descending trend, but

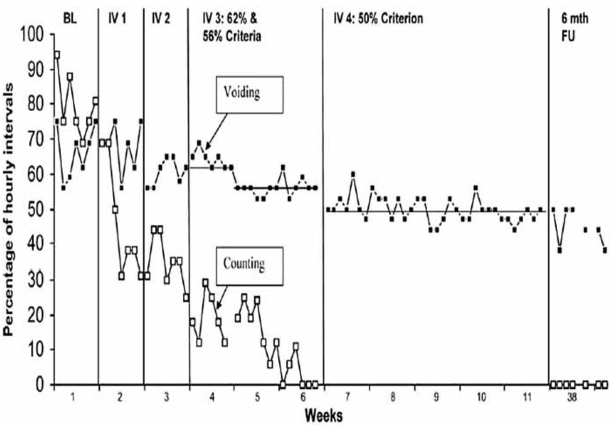


Figure 5. From Arco, L. (2008). Neurobehavioural treatment for obsessive-compulsive disorder in an adult with traumatic brain injury. *Neuropsychological Rehabilitation 18*(1), 109–124. Reprinted with permission. IRD = .86; Tau-U = .91; $p < .01$; Visual Analysis ES (S, M, L) MA = 0, 3, 5; PhD = 1, 2, 8.

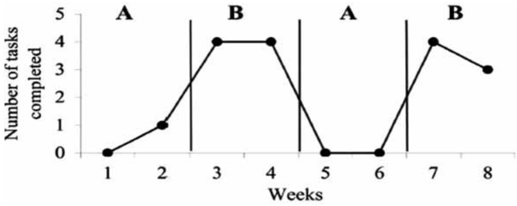


Figure 6. From McKerracher, G., Powell, T., & Oyebode, J. (2005). A single case experimental design comparing two memory notebook formats for a man with memory problems caused by traumatic brain injury. *Neuropsychological Rehabilitation, 15*(2), 115–128. Reprinted with permission. IRD = 1.0; Tau-U = 1.0, $p = .12$; Visual Analysis ES (S, M, L) MA = 1, 2, 5; PhD = 1, 3, 6.

also displays variable data and overlap with the baseline data. The effect size of .33 and .11 are both small and the *p* value is not significant. For visual analysis, nine students rated the effect size as small, followed by seven medium and three large, which reflects the influence that mean or trend lines can play in decision making, in both cases overemphasising behaviour change.

Figures 9 and 10 represent problems that occur with issues of scale. The calculated effect size for both IRD and TAU was 1.0, $p = .06$ and .04 level, respectively. Examining Figure 9, the visual analysis for all 19 students categorised this effect size as small. Interestingly, if you look back at earlier

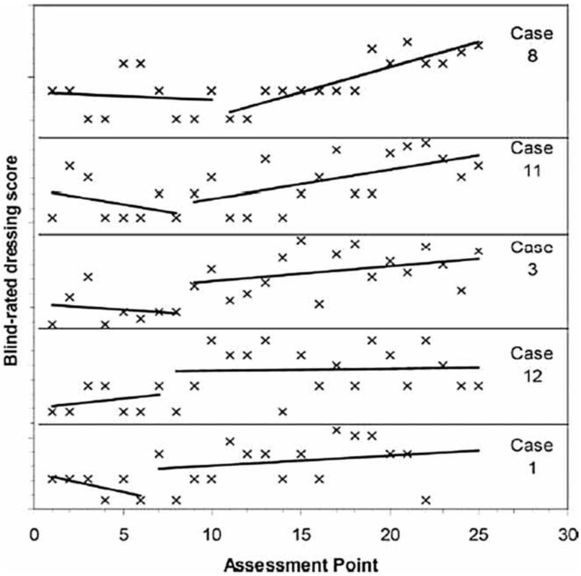


Figure 7. From Sunderland, A., Walker, C. M., & Walker, M. F. (2006). Action errors and dressing disability after stroke: An ecological approach to neuropsychological assessment and intervention. *Neuropsychological Rehabilitation, 16*(6), 666–683. Reprinted with permission. IRD = .12; Tau-U = .43, *p* = .07; Visual Analysis ES (S, M, L) MA = 2, 5, 1; PhD = 1, 6, 3.

figures you will see scales with variable ranges, some similar to this data series where most students ranked the effect as large. These inconsistencies due to scale are an important reason to incorporate effect sizes into any reporting of results. Figure 10 received more endorsement as a large effect, but fewer than half the students recognised this change. These graphs illustrate the importance of context. The statistical effect sizes are context free, whereas visual analysis is usually conducted with an understanding of what the data means (something our judges did not have). Thus, in scenarios like this where there are large effect sizes with no overlap, only knowledge of the context in which the data were collected will help one to determine the meaning of 100% nonoverlap.

Figure 11 (using the VAS3 line; VAS stands for Visual Analogue Scale and was a measure of positive mood) demonstrates a steadily decreasing trend for 3 of 4 baseline sessions (5, 3, 2) with a large jump on the fourth day (7). Intervention onset consists of the following data 7, 8, 8, 7, 6, 8.5, 8, 8.5. The decreasing baseline trend gives us one set of information, but the rise to intervention level effect in baseline presents another. How does one interpret this effect? Visual analysts rated this as small (5), medium (11), and large (3). These judges showed very little consistency, but a

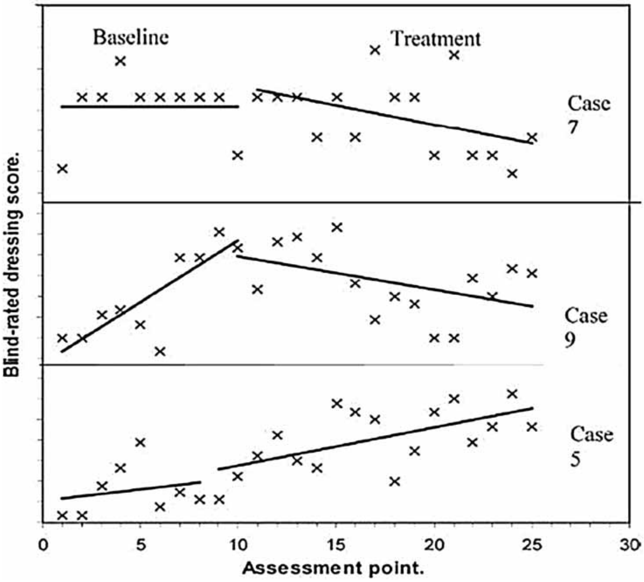


Figure 8. From Sunderland, A., Walker, C. M., & Walker, M. F. (2006). Action errors and dressing disability after stroke: An ecological approach to neuropsychological assessment and intervention. *Neuropsychological Rehabilitation, 16*(6), 666–683. Reprinted with permission. IRD = .33; Tau-U = .16; p = .51; Visual Analysis ES (S, M, L) MA = 5, 2, 1; PhD = 4, 5, 2.

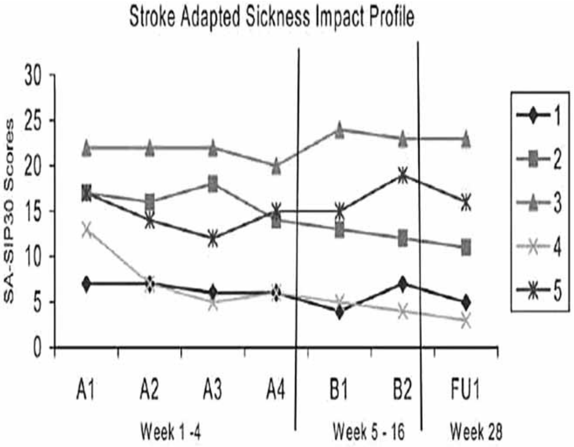


Figure 9. From Rasquin, S. M. C., Van de Sande, P., Praamstra, A. J., & van Heugten, C. M. (2008). Cognitive-behavioural intervention for depression after stroke: Five single case studies on effects and feasibility. *Neuropsychological Rehabilitation, 19*(2), 208–222. Reprinted with permission. IRD = 1; Tau-U = 1; p =.06; Visual Analysis ES (S, M, L) MA = 8, 0, 0; PhD = 11, 0, 0.

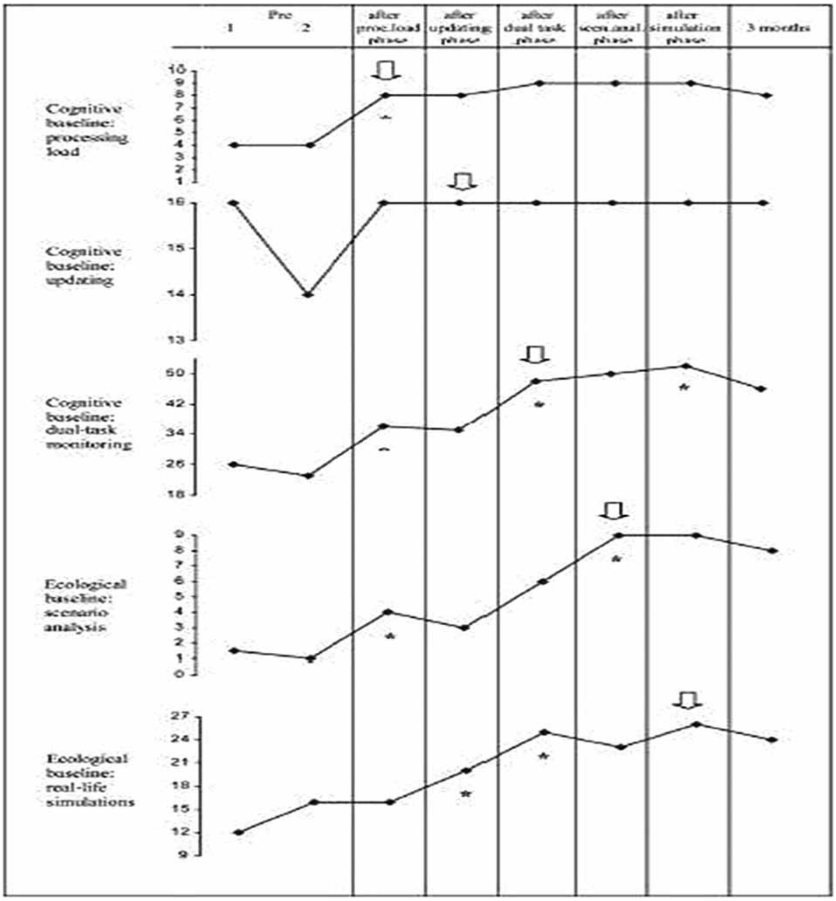


Figure 10. From Duval, J., Coyett, F., & Seron, X. (2008). Rehabilitation of the central executive component of working memory: A re-organization approach applied to a single case. *Neuropsychological Rehabilitation, 18*(4), 430–460. Reprinted with permission. IRD = 1; Tau-U = 1; p = .04; Visual Analysis ES (S, M, L) MA = 0, 6, 2; PhD = 2, 3, 7.

tendency towards medium. The IRD effect size of .75 and the Tau-U of .88 with a p value of .02 indicates that, statistically, this effect is large and significant.

Figure 12 demonstrates behaviour that does not change immediately with the onset of an intervention, likely because the dependent variable (Hair combing) required some development of skill that may not be immediate. A steady and large increase is followed by a clear deceleration and then another peak of improvement, and again a tapering off to resume baseline levels of behaviour. If this were a pretest–post-test design, there would be

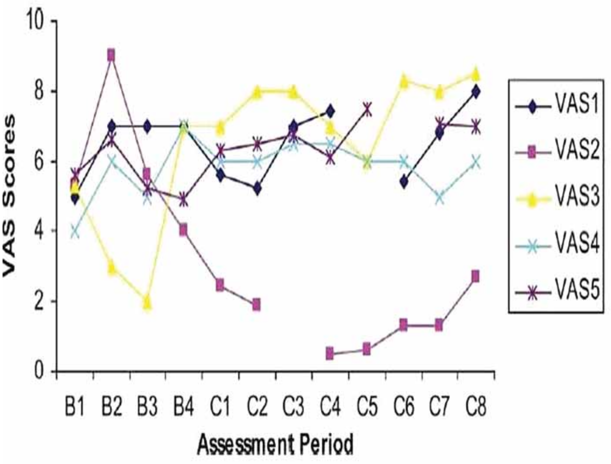


Figure 11. From Rasquin, S. M. C., Van de Sande, P., Praamstra, A. J., & van Heugten, C. M. (2008). Cognitive-behavioural intervention for depression after stroke: Five single case studies on effects and feasibility. *Neuropsychological Rehabilitation, 19*(2), 208–222. Reprinted with permission. IRD = .75; Tau-U = .88; $p < .02$; Visual Analysis ES (S, M, L) MA = 3, 5, 0; PhD = 2, 6, 3.

zero rates of behaviour at the beginning and zero at the end, demonstrating no change. The behaviour did change along a trajectory that could be captured with few points of data collection. Half of the visual analysts scored this as a large effect, yet the two effect size indices agree that this is more likely a moderate effect .78 and .57, yet it is statistically significant at $p < .01$.

ADDITIONAL EXAMPLES

Finally, in addition to demonstrating the use of two statistical methods (IRD and Tau-U) on 10 representative neuropsychological data sets and comparing those effect sizes with visual analysis, we identified four graphs to further the discussion of issues the single-case researcher faces when analysing data. The graphs are based on published data, but were modified for illustrative purposes.

Figure 13 results in a Tau-U = −.91, $p < .01$, and an IRD of .48. Ignoring the negative sign for Tau (which reflects the negative slope) there is a difference of .43, which is a large discrepancy between the two methods. In addition, this is an example where having lots of data points allows Tau-U to reach statistical significance, whereas the IRD of .48 suggests that there was a moderate amount of data overlap between baseline and treatment phases. In situations like this, it would be important to have a solid

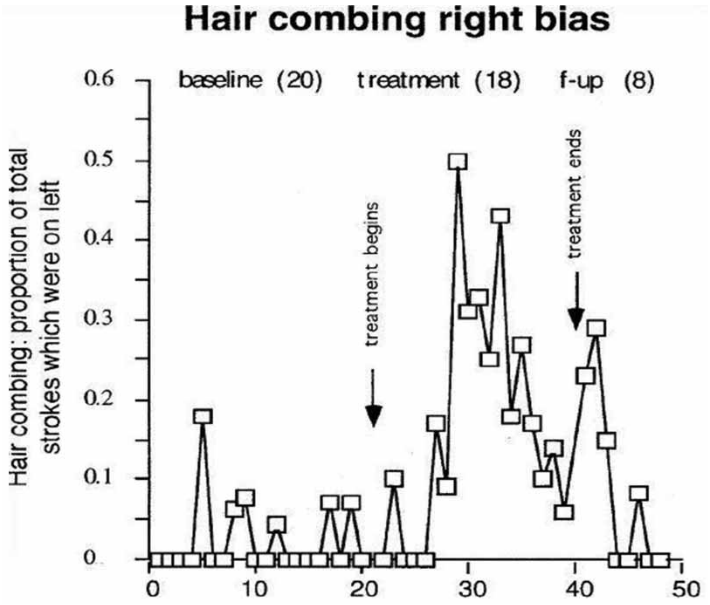


Figure 12. From Bergego, C., Azouvi, P., Deloche, G., Samuel, C., Louis-Dreyfus, A., Kashel, R., &Willmes, K. (1997). Rehabilitation of unilateral neglect: A controlled multiple-baseline-across-subjects trial using computerized training procedures. *Neuropsychological Rehabilitation, 7*(4), 279–293. Reprinted with permission. IRD = .78; Tau-U = .57; p < .01; Visual Analysis ES (S, M, L) MA = 2, 2, 4; PhD = 0, 3, 5.

understanding of these data and the conditions under which they were collected so one could determine the meaning of the change graphically displayed. It is unknown what the spike around session 78 means, but it appears that the treatment was having an effect. Only by factoring in the context of the study and how the data were collected and the variable measured, would one be able to determine if the IRD of .49 more accurately depicts the change (a fairly large degree of overlap) or if a statistically significant effect size based on Tau-U is more representative.

In Figure 14 there is a clear treatment effect with Tau-U = −1, p < .01, and IRD = 1. If one checks for baseline trend without first graphing these data, one may erroneously conclude that trend should be corrected for and report Tau-U = −.03, p = .79, when the baseline was corrected for trend. This serves as a reminder to always graph one's data and check the statistical results with visual analysis to make sure they make sense.

Examining Figure 15 suggests baseline trend may be problematic. Correcting for baseline trend results in Tau-U = .22, p = .33, whereas Tau-U with no baseline correction is .74, p = .02. IRD is .83, which is a difference of .44

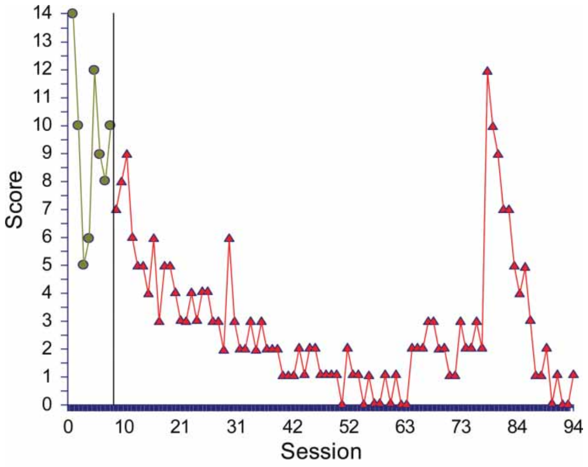


Figure 13. From O'Kearney, R. (1993). Additional considerations in the cognitive-behavioral treatment of obsessional ruminations: A case study. *Journal of Behavior Therapy & Experimental Psychiatry, 24*, 357–365.

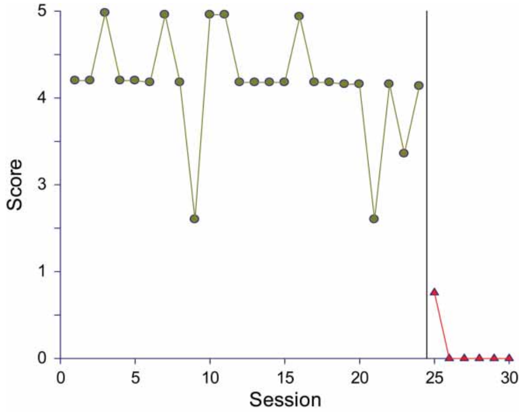


Figure 14. From Chadwick, P. (1994). Examining specific cognitive change in cognitive therapy for depression: A controlled case experiment. *Journal of Cognitive Psychotherapy*, *8*, 19–31.

after subtracting the baseline corrected Tau-U (.39). Although baseline trend is apparent, most clinicians would probably view the IRD of .83 as consistent with the treatment effect portrayed in Figure 15. If one corrects for baseline trend using Tau-U, then the results are non-significant, whereas the non-corrected Tau-U produces a statistically significant effect size. While Tau-U controls baseline trend in a conservative manner, some have reported tentative findings suggesting that the assumption that baseline trend continues

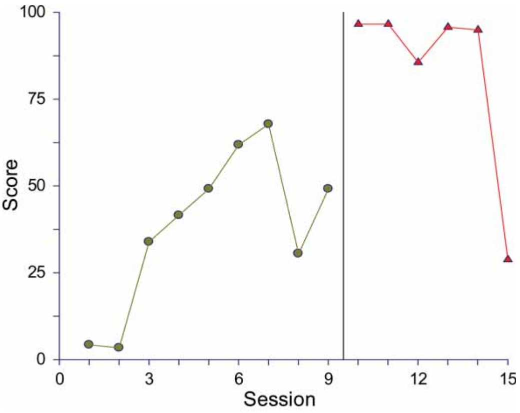


Figure 15. From Tollefson, N., Tracy, D. B., Johnsen, E.P., & Chatman, J. (1986). Teaching learning disabled students goal-implementation skills. *Psychology in the Schools, 23*, 194–204.

into the treatment phase should be questioned (Parker, Vannest, Davis, & Sauber, 2011b). Additionally, there is one point in the treatment phase that overlaps with the baseline data points. How much weight one gives the baseline trend and the final treatment data point may determine whether one views the above graph as showing at least a moderate treatment effect versus a small and non-significant one.

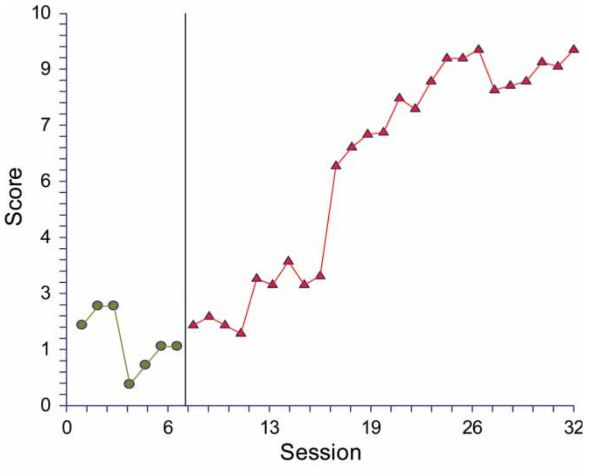


Figure 16. From Bujold, A., Ladouceur, R., Sylvain, C., & Boisvert, J.-M., (1994). Treatment of pathological gamblers: An experimental study. *Journal of Behavior Therapy & Experimental Psychiatry*, *25*, 275–282.

The graph in Figure 16 produced a Tau-U $= .87$, $p < .01$ and an IRD of .57, which is a difference of .30 between the two methods. Because IRD allows one to decide how data points are "removed" in its calculation, one can get different results. In this graph, if one "removes" the three data points in the baseline phase, the IRD is $25/25 - 3/7 = 1 - .43 = .57$. If one instead "removes" four data points from the treatment phase the IRD is $21/25 - 0/7 = .84 - 0 = .84$. Thus, this graph is a good example of the ability to manipulate IRD given phases with disproportional amounts of data. Parker et al. (2009) "removed" the smallest number of data points in their examples, but also state that "to the extent possible, data point removal should be balanced across the contrasted phases" (p. 141). Thus, how IRD is calculated may vary and no concrete rules have been developed for its use. In the example above, most researchers would probably concur that an IRD of .84 is an accurate effect size of the treatment effect portrayed visually.

DISCUSSION

Studies consistently indicate low agreement between visual and statistical results (Brossart et al., 2006; Jones, Weinrott, & Vaught, 1978; Park, Marascuilo, & Gaylord-Ross, 1990; Rojahn & Schulze, 1985). Most of these studies have methodological concerns (Brossart et al., 2006; Matyas & Greenwood, 1990) and the data from the present study demonstrate similar findings. Visual analysis is confounded by variable data, trend, and scale. Statistical analysis may be influenced by the same considerations so choices in analysis should be informed by the type of data (number of data points, presence of undesired trend in baseline or intervention, degree of overlap). Of the nonoverlap techniques presented here, IRD and Tau-U were compared across 10 illustrative graphs. First, these findings suggest that majority agreement between visual analysts and statistical analysis can be expected in cases of reasonably clear data even when visual analysts are not in perfect agreement. This holds true for small non-significant effects and large statistically significant effects. Second, we found that few data points can cause problems for visual analysts who may not be able to account for chance or probability of error in their assessment of a line graph. Third, we also found that the use of mean or trend lines typically assists in agreement between visual raters, but may not give "accurate" results in relationship to a statistical analysis of effect. Studies on the effects of trend lines suggest that even significant trends are unlikely to continue into the future. Specifically, the first five data points in a series are statistically unlikely to look like the second set of five (Parker, Vannest, Davis, & Sauber, 2011b). Fourth, another common problem in visual analysis is related to the issue of scale. Statistical analysis can provide clear empirical evidence of nonoverlap to assist in decision making when visual analysis can be misled. Even so, one must consider the context in which the data were collected and the nature of the variable being examined. Statistical methods can produce effect sizes, but they cannot factor in the multiple ways that context can impact the interpretation of one's data. Fifth, some data sets are prone to produce inconsistency in the visual analysis among multiple raters (e.g., Figures 8 and 11). Often these graphs have variability in the baseline or treatment phase that makes evaluation of the size of the treatment effect problematic. Some data sets will have a range of variability in both phases that may also prove troubling for visual analysts, but large amounts of variability are typically more of a challenge when combined with possible trend or just enough overlap to make the treatment effect appear somewhat ambiguous. In such situations, calculating one or more effect sizes can be an invaluable aid in evaluating the treatment effect.

Based on the data presented here and in previous studies, we recommend the following guidelines in selecting a statistical effect size analysis

1. Before conducting any data analysis conduct a design analysis, assess your design for functional relationships. This provides the foundation upon which interpretations of any effect sizes will be made.
2. Collect a sufficient amount of baseline data, particularly when trend plays a role or is expected. Short baselines reduce the ability of the visual analyst to interpret any treatment effect and increase the size of confidence intervals.
3. Trend needs to be corrected conservatively because the likelihood of trend continuation appears to be low. Trend is generally unreliable and unpredictable. The relationship of the length of data to the overcorrection of trend is inherent in most calculations so caution should be used.
4. Always conduct both visual analysis and statistical analysis, they should inform and reinforce each other.
5. Because all methods for calculating an effect size have limitations, the best practice may be to report multiple effect sizes. Reporting several effect sizes based on differing conceptual foundations (regression, dominance, overlap, etc.) will provide more information and should aid one in arriving at a more accurate interpretation. Reporting multiple fit indices is standard practice when one evaluates structural equation models. Each fit index has particular strengths, limitations, and meaning. Providing several fit indices aids in assessing model fit. Reporting several effect sizes of single-case data may also aid in interpreting treatment effects.

LIMITATIONS

This paper focused on one promising nonoverlap method (Tau-U) and issues related to integrating an effect size with visual analysis. Tau-U is a technique that is able to compare the baseline phase to the treatment phase, but it can also control for baseline trend and account for trend in the treatment phase. While this paper focused mostly on the comparison between baseline and treatment phases, the other variations of Tau-U are included in the output produced by the R syntax available online: https://dl.dropboxusercontent.com/u/2842869/Tau_U.R.

Inevitably, some problematic issues were not adequately discussed given the small number of examples presented. For instance, autocorrelation was not addressed and a limited number of illustrative problematic data sets were included. Only nonoverlap methods were addressed; there are many other statistical methods available for producing effect sizes. We also did not discuss how to apply phase A and phase B effect size methods to other design structures with multiple phases. This deserves additional research. Such research would be relevant for a large number of clinicians and researchers.

Some investigators, such as Wolery et al. (2010) and Haardorfer (2010), have argued that overlap measures do not produce effect sizes. Carter (2013) argues such notions are based on misconceptions about effect sizes, overlap methods do address the magnitude of effects, but they cannot speak to issues related to causality. Only a design analysis can address issues related to causality. It is also worth noting that group designs are focused on between-individual variation, whereas in single-case designs, the differences are based on within-individual variation. Thus, Carter (2013) notes that the within-individual variation would usually be more constrained than the variation seen between individuals and would produce relatively larger effect sizes.

CONCLUSIONS

The purpose of this article was to present justification for the use of effect sizes in reporting SCED results, to briefly review the steps for several nonparametric effect sizes, to present a variety of exemplars and compare the performance of IRD, Tau-U, and visual analysis, and to identify talking points and recommendations for using effect sizes in SCED.

Both relative risk ratios and dominance methods such as IRD and Tau-U have strengths to suggest their use, but neither is without limitations. The 2 x 2 table of IRD is easy to hand calculate, especially for short data series. Tau-U can also be hand calculated using a matrix. To calculate confidence intervals, *p* values, or to analyse longer data sets, statistical programs are likely to be necessary (e.g., https://dl.dropboxusercontent.com/u/2842869/Tau_U.R).

These non-parametric "bottom-up" approaches are readily understood and interpreted by interventionists. The logic in the analysis is consistent with visual analysis and the results are informative in the interpretation of treatment effects. Tau-U is particularly flexible given its ability to adjust for monotonic trend, it works accurately with few data points, and Tau-U can work with any type of design and any type of data. For example, there is no need for normally distributed data or interval data. If trend is present, it does not have to be a linear trend to be workable and the correction is moderate for either a baseline or an intervention phase. Tau-U has good statistical power of 91–115% of parametric tests (Parker, Vannest, Davis, & Sauber, 2011b). It also performs well in the presence of autocorrelation. Parker et al. (2011b) found that in 75% of the data sets they examined that had dangerous levels of autocorrelation, Tau-U values changed little after autocorrelation was removed (the remaining 25% showed larger changes in Tau-U values after autocorrelation was removed, but this consisted of less than 5% of the data sets examined). One limitation is that when 100% nonoverlap is obtained the method has reached its limit in assessing the treatment effect. For instance, imagine two data sets with 100% nonoverlap. One data set has the baseline and treatment phase data close together, but maintains 100% nonoverlap. The other has considerable distance between the baseline and the treatment phase. Both will produce an effect size of 1.0 but the *p* value may prove helpful in that the data set with minimal distance between phases will probably not achieve statistical significance. It should be noted that if one examines other variations of Tau-U, besides the phase A vs. B comparison, even with 100% overlap, the Tau-U value produced may be less than 1.0, especially if one controls for phase A trend. As a reminder, all statistical results must be interpreted in light of the context in which the data were collected. Only by considering the design analysis, the context in which the data were collected, the particular nature of the variable measured, conducting visual analysis of graphed data, and calculating one or more effect sizes based on different techniques, can one accurately evaluate the treatment effect.

REFERENCES

Barlow, D. H., & Nock, M. K. (2009). Why can't we be more idiographic in our research? *Perspectives on Psychological Science*, *4*, 19–21.

Barnett, S. D., Heinemann, A. W., Libin, A., Houts, A. C., Gassaway, J., Sen-Gupta, S., ... Brossart, D. F. (2012). Small *N* designs for rehabilitation research. *Journal of Rehabilitation Research & Development*, *49*, 175–186. doi: http://dx.doi.org/10.1682/JRRD.2010.12.0242

Blais, M. A., & Hilsenroth, M. J. (2007). Methodcentric reasoning and the empirically supported treatment debates. In S. G. Hofmann & J. Weinberger (Eds.), *The art and science of psychotherapy* (pp. 31–47). New York, NY: Routledge.

Brossart, D. F., Parker, R. I., Olson, E. A., & Mahadevan, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification*, *30*, 531–563. doi: 10.1177/0145445503261167

Carter, M. (2009). Effects of graphing conventions and response options on interpretation of small *n* graphs. *Educational Psychology*, *29*, 643–658. doi: 10.1080/01443410903204315

Carter, M. (2013). Reconsidering overlap-based measures for quantitative synthesis of single-subject data: What they tell us and what they don't. *Behavior Modification, 37*, 378–390. doi: 10.1177/0145445513476609

Cattell, R. B. (1988). The data box: Its ordering of total resources in terms of possible relational systems. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd Edn., pp. 69–130). New York, NY: Plenum Press.

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

Franklin, R. D., Allison, D. B., & Gorman, B. S. (Eds.) (1996). *Design and analysis of single-case research*. Mahwah, NJ: Lawrence Erlbaum Associates.

Greenspan, P., & Fisch, G. S. (1992). *Visual inspection of data: A statistical analysis of behavior*. Paper presented at the Proceedings of the Annual Meeting of the American Statistical Association, Alexandria, VA.

Grimmer, K., Bialocerkowski, A., Kumar, S., & Milanese, S. (2004). Implementing evidence in clinical practice: The 'therapies' dilemma. *Physiotherapy*, *90*, 189–194. doi: 10.1016/j.physio.2004.06.007

Haardorfer, R. (2010). Concerns with using Cohen's *d* and PND in single-case data analysis. *Focus on Autism and Other Developmental Disabilities*, *25*, 125–127. doi:10.1177/1088357610371274

Hagopian, L. P., Fisher, W. W., Thompson, R. H., & Owen-DeSchryver, J. (1997). Toward the development of structured criteria for interpretation of functional analysis data. *Journal of Applied Behavior Analysis*, *30*, 313–326.

Hojem, M. A., & Ottenbacker, K. J. (1988). Empirical investigation of visual-inspection versus trend-line analysis of single-subject data. *Journal of the American Physical Therapy Association*, *68*, 983–988.

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

Jones, R. R., Weinrott, M. R., & Vaught, R. S. (1978). Effects of serial dependency on the agreement between visual and statistical inference. *Journal of Applied Behavior Analysis*, *11*, 277–283.

Kahng, S. W., Chung, K.-M., Gutshall, K., Pitts, S. C., Kao, J., & Girolami, K. (2010). Consistent visual analysis of intrasubject data. *Journal of Applied Behavior Analysis*, *43*, 35–45.

Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd Edn.). New York, NY: Oxford University Press.

Kennedy, C. H. (2005). *Single-case designs for educational research*. Boston, MA: Allyn and Bacon.

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 What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Lamiell, J. T. (1981). Toward an idiothetic psychology of personality. *American Psychologist*, *36*, 276–289.

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. doi: 10.1177/0145445504272974

Manolov, R., & Solanas, A. (2009). Percentage of nonoverlapping corrected data. *Behavior Research Methods*, *41*, 1262–1271.

Manolov, R., Solanas, A., & Leiva, D. (2010). Comparing "Visual" effect size indices for single-case designs. *Methodology*, *6*, 49–58. doi: 10.1027/1614-2241/a000006

Manolov, R., Solanas, A., Sierra, V., & Evans, J. J. (2011). Choosing among techniques for quantifying single-case intervention effectiveness. *Behavior Therapy*, *42*, 533–545. doi: 10.1016/j.beth.2010.12.003

Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis*, *23*, 341–351.

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, *2*, 201–218.

Nesselroade, J. R. (1991). Interindividual differences in intraindividual change. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 92–105). Washington, DC: American Psychological Association.

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

Parker, R. I., & Brossart, D. F. (2003). Evaluating single-case research data: A comparison of seven statistical methods. *Behavior Therapy*, *34*, 189–211. doi: 10.1016/S0005-7894(03)80013-8

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

Parker, R. I., & Vannest, K. (2009). An improved effect size for single case research: Non-Overlap of All Pairs (NAP). *Behavior Therapy*, *40*, 357–367. doi: 10.1016/j.beth.2008.10.006

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. (2011a). Effect size in single-case research: A review of nine nonoverlap techniques. *Behavior Modification*, *35*, 303–322. doi: 10.1177/0145445511399147

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011b). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy*, *42*, 284–299. doi: 10.1177/0145445511399147

Robinson, O. C. (2011). The idiographic/nomothetic dichotomy: Tracing historical origins of contemporary confusions. *History & Philosophy of Psychology*, *13*, 32–39.

Rojahn, J., & Schulze, H. (1985). The linear regression line as a judgmental aid in the visual analysis of serially dependent A-B time-series data. *Journal of Psychopathology and Behavioral Assessment*, *7*, 191–206.

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.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin Company.

Shadish, W. R., Rindskopf, D. M., & Hedges, L. V. (2008). The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention*, *2*, 188–196.

Skiba, R., Deno, S., Marston, D., & Casey, A. (1989). Influence of trend estimation and subject familiarity on practitioners' judgements of intervention effectiveness. *Journal of Special Education*, *22*, 433–446.

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

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. doi: 10.1177/0022466908328009

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.