Exceptional Children

Vol. 75, No. 2, pp. 135-150.

The Improvement Rate Difference *for Single-Case Research*

RICHARD I. PARKER

KIMBERLY J. VANNEST
Texas A&M University at College Station

LEANNE BROWN
Cox Elementary School

ABSTRACT: *This article describes and field-tests the improvement rate difference (IRD), a new effect size for summarizing single-case research data. Termed "risk difference" in medical research, IRD expresses the difference in successful performance between baseline and intervention phases. IRD can be calculated from visual analysis of nonoverlapping data, and is easily explained to most educators. IRD entails few data assumptions and has confidence intervals. The article applies IRD to 166 published data series, correlates results with three other effect sizes:* R^2*, Kruskal-Wallis W, and percent of nonoverlapping data (PND), and reports interrater reliability of the IRD hand scoring. The major finding is that IRD is a promising effect size for single-case research.*

Integral to applied behavior analysis (ABA) is the frequent measurement of client behavior over time, and its graphic display to guide decisions for managing interventions (Baer, Wolf & Risley, 1968; Sidman, 1960; Skinner, 1938). Inferences about the cause and amount of behavioral change are made from visual analysis of graphed data, to detect differences of "sufficient magnitude to be apparent to the eye" (Parsonson & Baer, 1978). However, when data have high variability or "bounce," client improvement has proved difficult to judge, requiring visual heuristics such as trend lines (Cooper, Heron, & Heward, 1987). Another adjunct to visual analysis is to draw "envelopes" around trend lines to indicate their stability (Lovitt, 1977; White & Haring, 1980). Thus, the best visual analyses commonly are supported by simple statistics-based heuristics.

Parsonson and Baer (1978) eschewed statistical analyses of ABA data, stating that an advantage of visual analysis alone is its conservatism; only visually prominent effects are accepted. However, published single-case research studies do not always portray unequivocally large results. Glass (1997) reviewed several volumes of the *Journal of Applied Behavior Analysis* and found frequent "small, weak and ephemeral effects, or effects that are entirely illusory" (p. 598). A recent examination and re-analysis of over 150 published single-case research data series found less than half with large effects (according to both visual and statistical analyses), and more than one

quarter showed small or debatable results (Parker, Cryer, & Byrns, 2006). Yet in the face of these modest results, only visual analysis has been employed in 90% of published single-case research studies, this rate unchanged over 3 decades (Busk & Marascuilo, 1992; Kratochwill & Brody, 1978; Parker & Brossart, 2003).

The prevalence of published small and moderate-size single-case research results, and the need to unambiguously describe these results, has led some to use statistics, to a limited extent. For decades, researchers have drawn trend lines and computed slopes to describe "rate of improvement." Percentile-based "envelopes" around the trend lines describe slope variability (Lindsley, 1971; Lovitt, 1977; White, 1986; White & Haring, 1980); percent of nonoverlapping data (PND; Scruggs, Mastropieri, & Casto, 1987) suggests size of experimental effect. Although the "logic of decision making that underlies single-case investigations is compatible with statistical reasoning" (White, Rusch, Kazdin, & Hartmann, 1989, p. 283), statistics for single-case research are still in an early stage of development.

The statistical summary with fastest rising popularity across research models is the *magnitude of effect* or *effect size* index (Kirk, 1996; Thompson, 2002a; Wilkinson & The Task Force on Statistical Inference, TFSI, 1999). American Psychological Association (APA) publication standards (2001) now require effect sizes for almost all quantitative research. Though not yet well accepted in single-case research, effect size has been described as the "obvious choice" for summarizing single-case study effects (Busk & Serlin, 1992, p. 192). However, an effect size cannot duplicate the breadth and integrated nature of holistic visual analysis (e.g., simultaneous consideration of mean or median level shifts, trends and trend differences, precipitous behavior changes at intervention onset, data variability within and across phases, differences in trend line intercepts at intervention points, curvilinear progress, lag or delay in response to intervention, etc.). The effect size does not compete with this breadth, but rather serves a narrower purpose, that of quantifying the amount of behavioral change between contrasted phases.

An effect size alone cannot communicate whether the intervention caused the improvement. "Statistics do not demonstrate causality. You need a rigorous design for that " (Bloom, Fischer, & Orme, 1999, p. 572). Nor can an effect size alone tell what type of improvement was measured: whether in trend, in mean or median level, or both. Nor can one effect size necessarily summarize a full single-case research design; some complex designs may require multiple phase contrasts for adequate summary. Despite these limitations, we contend that an effect size is a useful supplement to visual analysis, especially when information about the phase contrast and design context are also well described.

Effect sizes are simply a standardized expression of the amount of behavior change between phases. They do not address practical significance or clinical significance, values which can be overlaid on effect sizes, as a second step. Practical significance is achieved when behavior change has a practical impact on the client's daily life (Shaver, 1991; Thompson, 2002b), and the more demanding clinical significance is change from a dysfunctional to functional range of activity (Jacobson, Follette, & Revenstorf, 1984). However, effect sizes are a beginning point for overlaying social value judgments by teachers, school psychologists, and clinicians.

An effect size is a useful supplement to visual analysis, especially when information about the phase contrast and design context are also well described.

Effect size calculation can serve the primary goal of establishing a functional relationship between intervention and behavior. To establish this relationship, Horner et al. (2005) recommend examining a minimum of three phase shifts, as in an ABAB design or a three-series multiple baseline design. Visual analysis skills illuminate behavior change at each of these phase shifts, attending to data variability and, especially, to data trend. After examining the intercept gap for each phase shift, the researcher is able to answer the question of a functional relationship. The follow-up questions of (a) degree of change, and (b) reliability of the change can be answered by calculating an effect size for each of the three (or more) phase shifts, in turn. If a functional relationship was evidenced,

then an overall (omnibus) effect size can be conducted on the entire design. The omnibus analysis includes all component phase shifts to produce a single effect size with greater reliability (smaller confidence interval) than those of the component effect sizes.

Combining effect sizes with visual analysis offers at least four advantages to single-case research: objectivity, precision, certainty, and general acceptability. The objectivity of an effect size is useful when subjective visual judgments disagree. Adequate agreement by visual judges continues to be a challenge in the single-case research field, even among experienced researchers (Brossart, Parker, Olson, & Mahadevan, 2005; Harbst, Ottenbacher, & Harris, 1991; Ottenbacher, 1990; Park, Marascuilo, & Gaylord-Ross, 1990). Objectivity also is an asset when communicating results unambiguously to a new audience of stakeholders (Huitema, 1986).

Precision is a second advantage of including an effect size with visual analysis. The visual analysis reliability studies (Harbst et al., 1991; Ottenbacher, 1990; Park et al., 1990) consistently indicate that only gross judgments can be made accurately and consistently. If only large effects were of interest and published, as per the recommendation of Parsonson and Baer (1992), then visual analysis may be adequate. However, this does not describe current practice, as many published single-case research studies show only moderate- and small-size results (Glass, 1997). In a recent survey of over 150 published single-case research datasets, 25% showed small, debatable, or negligible results, according to both visual and statistical analysis (Parker et al., 2006).

Certainty is a third advantage of including effect sizes with visual analysis. Certainty is the level of confidence one may have that results are not due to chance alone. In a short dataset, a large effect may be nullified or reversed by only a few additional data points. Though the measured effect size may be large, we can have little confidence in it. Level of confidence is indicated by a confidence interval bracketing the effect size. Confidence intervals are available for all standard effect sizes (Cohen's *d*, Hedges's *g*, *R*, R^2, phi, *W*, odds ratio, risk difference) because they all have known sampling distributions. However, for PND (Scruggs et al., 1987) the sampling distribution is unknown.

The fourth advantage of visual plus statistical analysis is acceptability by the broader research field. ABA was developed, practiced, and published in relative isolation from other social science research. However, broad changes in education and psychology research related to evidence-based practices, interventions, or treatments are bringing single-case research under the scrutiny of a broader professional community. For identifying successful evidence-based treatments through meta-analysis, effect sizes are a necessity (Forness, 2001; Mostert, 2001), and single-case research's lack of standard effect sizes is excluding these studies from meta-analyses (Allison & Gorman, 1993; Giles, 1990; Strube, Gardner, & Hartmann, 1985).

Guidelines for single-case research are being strengthened by such fields as school psychology (Kratochwill & Stoiber, 2002); special education (Horner et al., 2005); and clinical psychology (Chambless & Ollendick, 2001). These changes should help gain greater acceptability for single-case research by the broader research community. Finally, single-case research is being scrutinized at the federal level, where more stringent design and analysis standards have been set by the new Institute of Education Sciences (IES, 2005).

Of the dozens of established effect sizes (Cohen, 1988; Kirk, 1996), most pose problems for applied researchers, including: (a) having opaque or esoteric meanings, (b) assuming data properties lacking in most single-case research datasets, (c) being technically difficult to produce, (d) being difficult to meld with visual analysis, and (e) encouraging oversimplified misinterpretations. A prime example of esoteric meaning is R^2 as percent of variance accounted for, which presumes understanding of ordinary least-squares variance partitioning. Similarly, the interpretation of Cohen's *d* as standardized mean difference presumes understanding of normal distributions, standard deviation units, and *Z* scores. Regression R^2 and ANOVA or *t* test (or hand-calculated) Cohen's *d* are statistically equivalent (Rosenthal, 1991; Wolf, 1986). Many educators lack the background for understanding either of these common effect sizes (May, 2004; Rosenthal, Rosnow, & Rubin, 2000; Weiss & Bucuvalas, 1980).

The second limitation of effect sizes such as *R*, R^2, Cohen's *d* and Hedges's *g* is that they assume that data are serially independent, normally distributed, and have constant variance. Standardized mean difference effect sizes (Cohen's *d*, Hedges's *g*) are statistically equivalent to regression effect sizes (*R*, R^2), and have the same data assumptions. Most single-case research data fail to meet at least one of these assumptions (Matyas & Greenwood, 1996; Parker & Brossart, 2003). A recent examination of 166 published multiple baseline designs found over two thirds failing to meet either normality or equal variance assumptions, and another two thirds were undesirably autocorrelated (Parker, 2006).

The third limitation is the difficulty of carrying out the statistical analyses. Despite the ease of menu-driven statistics, interpretating output may require statistical knowledge beyond introductory coursework. For parametric analyses, the user must be able to interpret tests for normality and homogeneity of variance. This is especially difficult with short data series, for which the best tests of assumptions are invalid and more subtle visual analysis of residual scores is necessary (Hintze, 2007). In calculating Cohen's *d*, formulas must be adjusted due to non-constant variance and unbalanced phase lengths (Busk & Serlin, 1992).

A fourth limitation is that most effect sizes do not blend well with visual analysis (Parsonson & Baer, 1992). Over the history of ABA, visual analysis has developed its own heuristics, which do not include effect sizes. Visual analysis typically focuses on nonoverlapping data, spatial separation of score clusters, estimated proportion of scores above or below performance standards, estimated data trends and changes in those trends, and discontinuity in trend lines at intervention onset—as well as noting precipitous changes in individual data points around intervention onset times. These heuristics entail no more than ordinal-level data assumptions, not the more stringent assumptions of R^2 and Cohen's *d*.

Finally, effect sizes presented alone, without context, are likely to be misinterpreted (Thompson, 2006). An effect size needs to be accompanied by contextual details of the design, the intervention, the particular phase contrast employed, and the type of statistical analysis (Durlak, 2002; Kirk, 1996; Rosnow & Rosenthal, 1989; Wilkinson & TFSI, 1999). Especially in multiseries and complex multiphase designs, the consumer needs to know which phases were contrasted and excluded from effect size calculations. The consumer also needs to know what statistical model was used, because effect size magnitudes vary by model (Parker & Brossart, 2003). In short, an effect size must be presented with guidelines for its interpretation.

The purpose of this article is to introduce the *improvement rate difference* (IRD) for single-case research data, an effect size with advantages over standardized mean difference (Cohen's *d*, Hedges *g*) and variance accounted for (*R*, R^2) alternatives. IRD's advantages include (a) accessible interpretation as the difference in improvement rates between baseline and treatment phases; (b) simple hand-calculation; (c) compatibility with PND from visual analysis; (d) known sampling distribution, so confidence intervals are available; (e) proven track record (as *risk difference*) in hundreds of evidence-based medical research studies; (f) few data distribution assumptions; and (g) application to complex single-case research designs and multiple data series.

IRD is defined as the improvement rate (IR) of the treatment phase(s) minus the improvement rate of the baseline phase(s): $IR_T - IR_B = IRD$ (Cochrane Collaboration, 2006; Sackett, Richardson, Rosenberg, & Haynes, 1997). Its calculation is described in the Method section. IRD has a solid record of use in evidence-based medicine, under the name of *risk reduction* or *risk difference*; renaming it in this article reflects the focus in single-case research on client improvement rather than reduction of risk for disease or death. Risk difference is valued by medical researchers for its interpretability, for the fact that it does not require unwarranted data assumptions, and because it has easily obtained confidence intervals (CIs; Altman, 1999; Sackett et al., 1997). The prestigious Cochrane Collaboration (2006) promoted risk difference as a summary of treatment efficacy for evidence-based medicine, and has helped initiate its counterpart for educational research, the Campbell Collaboration (Petrosino, Boruch, Rounding, McDonald, & Chalmers, 2000; Wolf, 2000). The Campbell Collaboration has spurred higher federal standards for funded educational

research, including stronger designs and effect sizes with CIs (Whitehurst, 2004).

IRD is closely related to the percent of all nonoverlapping data (PAND; Parker, Hagan-Burke & Vannest, 2007). Parker and colleagues described PAND as an overlap statistic which could be converted to the established effect size, phi. They stated that in a balanced, symmetrical 2 × 2 table, phi could be recalculated as the difference between two proportions, and proportions statistics used, rather than chi-square. That extension has been carried out in this article and applied to a sizeable sample, resulting in an IRD index bearing strong correlation to PAND and phi. Both phi and IRD are considered strong, useful effect sizes for single-case research. Our field test of IRD to 166 simple AB contrasts from published single-case research studies helps answer questions practitioners might have about IRD:

1. Can IRD be reliably calculated?
2. How do IRD results relate to better known effect sizes?
3. What effect size magnitudes are typically found in IRD?
4. How well does IRD discriminate among single case datasets?

METHOD

IRD is calculated as the difference between two IRs (Cochrane Collaboration, 2006; Sackett et al., 1997). The IR for each phase is defined as the number of "improved data points" divided by the total data points in that phase:

$$\frac{\text{\# impr.}}{\text{\# total}} = \text{IR}$$

An improved data point in baseline is defined as one that ties or exceeds any data point in the treatment phase. An improved data point in the treatment phase is defined as any which exceeds all data points in the baseline phase. "Exceeds" refers to higher levels of behaviors we wish to increase (e.g., homework completion) and to lower levels of behaviors we wish to decrease (e.g., tantrums). Improved data points are identified visually; IRD is calculated as the difference between two independent proportions.

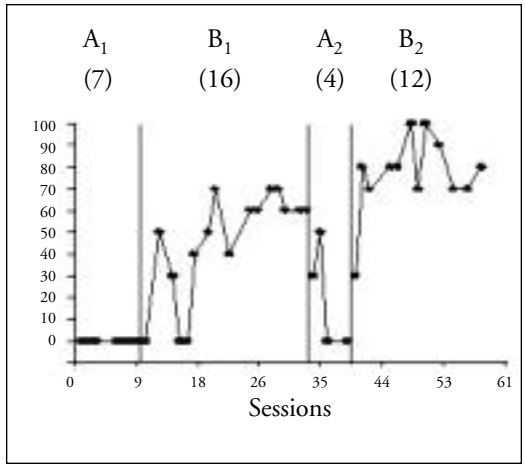
IRD is here modified from its use in group medical designs to better suit single case designs: The control group or condition becomes the baseline phase(s), and the treatment group or condition becomes the intervention phase(s). Improved (or successful) versus unimproved (or unsuccessful) data are defined by data overlap between phases. In the baseline phase, unimproved data do not overlap (equal or exceed) any treatment phase data, unlike improved baseline data. In the treatment phase, unimproved data equal or fall below one or more baseline data points. Overlapping data (for an AB contrast) are defined as the fewest data points that would have to be removed (from either phase A or B) to eliminate all data overlap between phases.

The maximum IRD score is 100% or 1.00, in which case all intervention phase scores exceed all baseline scores (in an improvement direction). IRD is calculated as the difference in these phase-specific improvement rates: 100% (Phase B) – 0% (Phase A) = 100%. If the *N*s for baseline and intervention phases were 12 and 45, respectively, the calculation would be 45/45 – 0/12 = 45/45 = 1.00. An IRD of 50% (.50) indicates that half of the scores are overlapping, so did not improve from Phase A to B. When IRD = .50, there is only chance-level improvement from baseline to treatment phases. A negative IRD score is possible, indicating deterioration below baseline levels.

The confidence one can have in an obtained IRD is defined by its confidence interval (CI), which brackets the IRD. Very wide CIs indicate that the computed IRD is not very trustworthy, regardless of its size. Width of CIs can also be interpreted as measurement precision, with narrow CIs showing more precision (Harper, 1999). Most statistical packages (e.g., SPSS, SAS, NCSS, S-Plus) provide CIs for the difference between two proportions, under "proportion statistics" or "risk analysis." This study used NCSS version 2006, which offers several CI options. Besides being included in standard educational research statistical packages, CIs are also available in less-known software. StatsDirect (version 2.5.5; Buchan, 2006), a program designed by and for medical researchers, graphically displays risk difference results with CIs. Specialized meta-analysis software such as MetaWin (Rosenberg, Adams, & Gurevitch,

FIGURE 1

Improved Verbal Responding From Using a Lag Reinforcement Schedule With a Child With Autism, ABAB Design



2000) and dr-ROC (Mitchell, 2005) also calculate risk differences with CIs. In addition, WinPEPI software for epidemiologists (Llorca, 2002) is freely available from a biomedical Web site (http://www.epi-perspectives.com/ content/1/1/6), and Ian Buchan, the author of StatsDirect (2006), provides free interactive Web-based calculations from the University of Manchester Medical School (http://www.phsim.man.ac.uk/).

APPLIED EXAMPLES

We applied the IRD procedure to two published datasets, purposefully selected for their ABAB and multiple baseline designs. We selected datasets with visually apparent effective treatments yet with some data overlap between phases. The examples are not intended to present model designs or results, but rather typical published data that could benefit from quantitative analysis. The first example, an ABAB reversal design (Lee, McComas, & Jawor, 2002), highlights 1 participant from a broader design entailing 3 participants. Lee and colleagues used a lag reinforcement schedule as an intervention to increase varied and appropriate verbal responding by three 7-year-old boys with autism. The baseline condition utilized differential reinforcement without a lag schedule. According to the Horner et al. (2005) "three phase shifts" guideline, the ABAB design qualifies as an experimental design. Figure 1 provides the data count for each phase in parentheses.

The three critical contrasts in the Lee et al. (2002) design are A_1 versus B_1, B_1 versus A_2, and A_2 versus B_2. The phase contrast that best reflects the full design is A_1A_2 versus B_1B_2. Figure 2 artificially segments the ABAB design to clarify IRD calculations for the first three contrasts.

For each contrast we asked "What is the smallest number of data points needing removal from either phase to eliminate all overlap between the two phases?" Data points may be removed

FIGURE 2

Three Strategic Contrasts in the ABAB Design Study

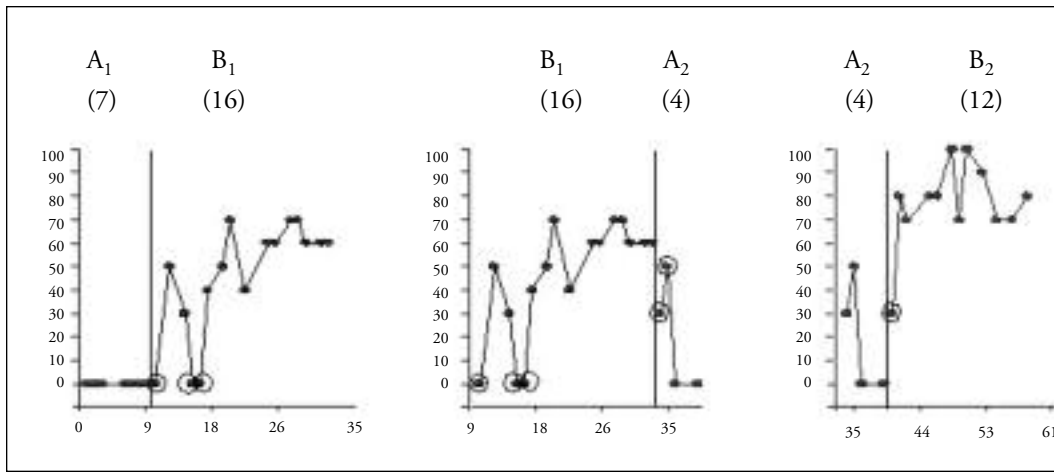


TABLE 1

Results of Three IRD Visual Analyses of an ABAB Reversal Design

	Condition		Condition		Condition	
Improvement	*Baseline Phase A_1*	*Treatment Phase B_1*	*Baseline Phase B_1*	*Treatment Phase A_2*	*Baseline Phase A_2*	*Treatment Phase B_2*
Improved	0	13	13	2	0	11
Not improved	7	3	3	2	4	1
Totals	7	16	16	4	4	12

from Phase A, Phase B, or both. Data points also are defined as overlapping if they have the same value across two phases; a data point "needing removal" is classified as "improved" if from a baseline phase, and as "not improved" if from a treatment phase (see Table 1). For the A_1 versus B_1 contrast, the smallest number needing removal was three data points (circled in Figure 3); for the B_1 versus A_2, contrast, the fewest needing removal was five; for the A_2 versus B_2 contrast, the fewest was one. There may be more than one equally good solution to the smallest number of data points needing removal, so, to the extent possible, data point removal should be balanced across the contrasted phases.

We used Table 1's interior and total cell values to calculate IRD. For the A_1 versus B_1 contrast, the improvement rate for the baseline phase was 0/7 = 0%; for the treatment phase it was 13/16 = 81%. Their difference, the IRD, was 81% – 0% = 81%. For the second contrast, B_1 versus A_2, the improvement rates were baseline: 2/4 = 50%; treatment: 13/16 = 81%. Their IRD was 81% – 50% = 31%. The third, A_2 versus B_2, contrast's improvement rates were baseline: 0/4 = 0% and treatment: 11/12 = 92%, for an IRD of 92% – 0% = 92%.

The three IRD values (81%, 31%, 92%) may be averaged together for a full design or omnibus IRD of 68%. An alternate method for an omnibus IRD is to conduct an A_1A_2 versus B_1B_2 contrast, by visually scanning for any Phase A data overlapping with any Phase B data. The same question is asked: "What is the smallest number of data points which could be removed to eliminate all overlap between the A and B phases?" Figure 3 provides a solution, summarized in Table 2.

The IR for the two baseline phases was 2/11 = 18%, and 25/28 = 89% for the treatment phases; the IRD was 89% – 18% = 71%, not far from the averaged 68% IRD obtained above.

An effect size alone tends to give readers a false sense of precision. The confidence we should have in an obtained IRD depends largely upon the amount of data it is based on and its magnitude. Both an IRD calculated from a small number of data points and a small IRD warrant little confidence. The precision of an IRD is indicated by the

FIGURE 3

Data Points Needing Removal to Eliminate Overlap for an A_1A_2 Versus B_1B_2 Contrast

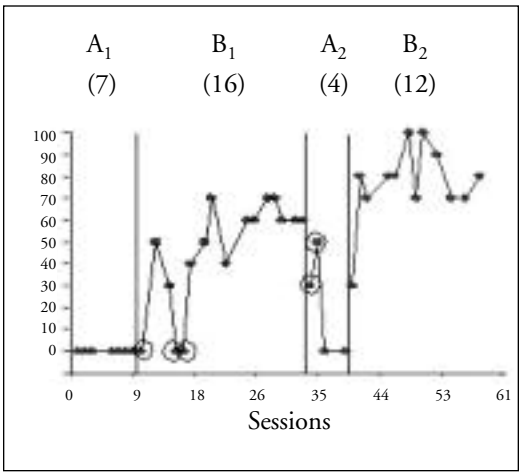


TABLE 2

Summary of Visual Analysis of an A_1A_2 Versus B_1B_2 Contrast

	Condition	
Improvement	*Baseline Phase A_1A_2*	*Treatment Phase B_1B_2*
Improved	2	25
Not improved	9	3
Totals	11	28

TABLE 3

Summary Results for Four Contrasts Conducted With IRD

Contrast	*Input: Two Proportions*	*Output: IRD With 85% CI*
A_1 versus B_1	0/7, 13/3	.69<<.81>>.94
B_1 versus A_2	13/3, 2/2	–.06<<.31>>.69
A_2 versus B_2	0/4, 11/1	.83<<.92>>1.00
A_1A_2 versus B_1B_2	2/9, 25/3	.53<<.71>>.91

Note. IRD = improvement rate difference; CI = confidence interval.

CIs which bracket it, forming upper and lower limits. These CIs are provided as standard output from a test of two proportions. Among the most reliable CIs are those based on bootstrapping, which sidesteps the problem of odd or asymmetrical data distribution shapes. We prefer 85% or 90% CIs for most clinical decision-making, though 90% or 95% are preferred for publishing.

We obtained CIs for an IRD from the NCSS "two proportions" test module. Table 3 summarizes the two proportions input and CI output for each of the four IRD values calculated thus far: 81%, 31%, 92%, 71%.

Obtaining CIs for effect sizes in single-case research can be humbling. For these examples, our level of certainty for clinical uses (85%) extends .10 to .20 points or more above and below the obtained IRD. The CIs are most optimistic for the A_2 versus B_2 contrast, which yielded the largest IRD (.92). The discouraging CIs for the B_1 versus A_2 enclose zero, so we cannot say with 85% certainty that for the obtained IRD of 31%, the true IRD is different from zero.

Our second example of IRD application is a multiple baseline design by Kennedy, Cushing, and Itkonen (1997), which measured the increase in social contacts by students with severe disabilities within general education classes. The intervention was systematic social support of individuals in the mainstreamed classrooms. One design used was a multiple baseline across two classrooms, Classes 2 and 6 during the school day, offering less design strength or internal validity than three or more baselines. Figure 4 presents the data for one dependent measure, the number of peer contacts made, with numbers of data points per phase in parentheses.

The contrasts which make most sense are A versus B for each baseline: A_1 versus B_1 for Class 2, and A_2 versus B_2 for Class 6. Once again, for each contrast we asked the question: "What is the minimum number of data points to be removed to eliminate all overlap between the contrasted phases?" If there are two equally good solutions, choose the one that removes a similar number of data points from each phase. For Class 2, Figure 3 shows (in circles) a "best solution": removing three data points. The same number need to be removed from the Class 6 data to eliminate all overlap.

Table 4 presents the results of visual analysis of data overlap. As in the first example's dataset, we calculated the improvement rates from the information in the table. For Class 2, the baseline phase improvement rate was 1/8 = 13%, and the treatment phase was 25/27 = 93%; the IRD was 93% – 13% = 80%. The test of two proportions (1/7, 25/2) gives the 85% bootstrap CI as: .64<<.80>>1.00. For Class 6, the baseline phase was 1/18 = 6%; treatment phase was 15/17 = 88%; and IRD was 88% – 6% = 82%. The 85% bootstrap CI for the two proportions (1/17, 15/2) was: .71<<.82>>.99. Both CIs are narrow enough

FIGURE 4

Increased Social Contacts by a Student With Severe Disabilities Over Two Environments, Multiple Baseline Design

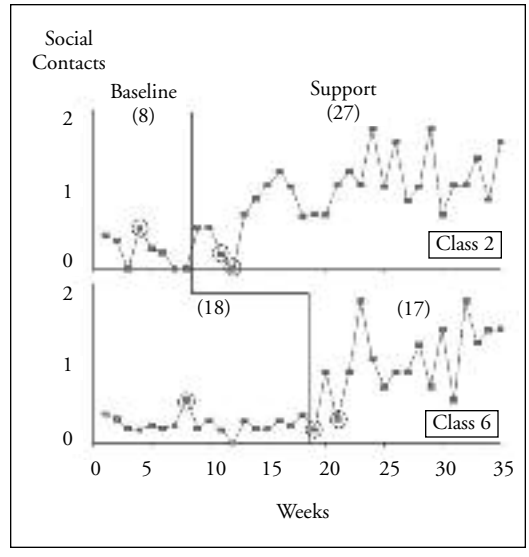


TABLE 4

Results of Visual Analysis of Multiple Baseline Series Design Study

Improvement	*Condition (Class 2)*		*Condition (Class 6)*	
	Baseline Phase A_1	*Treatment Phase B_1*	*Baseline Phase A_2*	*Treatment Phase B_2*
Improved	1	25	1	15
Not improved	7	2	17	2
Totals	8	27	18	17

to give us reasonable confidence in the obtained IRD values.

We obtained an IRD for the entire design by averaging the two individual IRD values (.80, .82) to obtain .81 or 81%. Alternatively, the two baselines in Table 4 can be added together, and an omnibus IRD calculated, providing CIs. For the omnibus IRD, the baseline IR was 2/26 = .08, the treatment IR was 40/44 = 91%, and the IRD was 91% – 8% = 83%. The 85% bootstrap CI for the two proportions (2/24, 40/4) was .75<<.83>>.93, more precision (narrower CI width) due to the increased number of data points. All of these calculations were conducted on numbers rounded to two decimal places for ease of replication, which could cause results here to differ from exact results by .01 to .02.

IRD FIELD APPLICATION

Because IRD is new to single-case research, we applied it to 166 A-versus-B contrasts from published data series (list of articles available upon request). We conducted only simple A-versus-B contrasts to permit comparability among the sample datasets and replicability by other researchers.

The field test included a check of the interrater reliability of the visual judgments required by the IRD procedure. In addition, we compared IRD with the better known parametric Pearson R^2 and the rank-based Kruskal-Wallis test (with its effect size, *W*), as well as Scruggs et al.'s (1987) PND. We also examined the distribution of IRD scores to identify typical values for published single-case research studies.

We obtained datasets by digitizing published graphs from ERIC and PsychLIT literature searches. We used multiple search terms to obtain single case design articles published over the past 25 years. This search generated a convenience sample, including several design types. We digitized all graphs with easily distinguishable individual data points using I-Extractor software version 1.0 (Linden Software Ltd, 1998). We included in this study only initial A-versus-B contrasts from the scanned datasets. The literature search and digitizing procedure are described in greater detail in previous articles (Parker & Brossart, 2003; Parker et al. 2005).

From 67 published articles, we obtained 166 initial A-versus-B contrasts. The average AB series contained 19 data points, 8 in baseline phase and 9 in intervention phase. The interquartile range (middle 50% of the series) was 14 to 24 data points, and the middle 90% ranged 10 to 30 data points. We analyzed the 166 contrasts by IRD, as well as by Pearson's R^2, the nonparametric Kruskal-Wallis *W*, and the Scruggs et al. (1987) PND.

RESULTS

Rater Reliability

We trained two graduate students in educational psychology/special education in the IRD procedure, who then applied it to the 166 datasets. The two sets of scores correlated $R = .96$ with one another. Seventy-two percent of the ratings were identical between the two raters. Differences above IRD = .10 were found in only 13% of the ratings. This was the first attempt by the raters to apply IRD to published data.

TABLE 5

Intercorrelations Among IRD, R^2, Kruskal Wallis W, and PND, Based on 166 Published AB Contrasts

	IRD	R^2	*K-W W*
R^2	.856		
Kruskal-Wallis W	.861	.920	
PND	.826	.746	.748

Note. IRD = improvement rate difference; K-W W = Kruskal-Wallis W; PND = percent of nonoverlapping data.

Intercorrelations

Any new measure such as IRD should be validated by existing measures with a history of use in the field. The three comparison measures employed here include the one with most statistical power (Pearson's R^2), the most powerful nonparametric technique available (Kruskal-Wallis *W*), and the most-used single-case research measure (PND). Table 5 presents intercorrelations among the four measures.

IRD showed high-moderate correlation (.86) with the established effect sizes R^2 and Kruskal-Wallis *W*, indicating validation support. IRD also was substantially related (.83) to the established overlap index, PND (Scruggs et al., 1987). Among the three external criteria (R^2, *W*, and PND), the first two were closely related (.92), though the first is based on explained variance and the second on rank order. The third external measure, PND, was moderately related (.75) to the other two.

Discriminability

A comparison between two or more uniform probability distributions can reveal their relative strengths and weaknesses at discriminating among single-case research contrasts with a variety of effect sizes (Cleveland, 1985). A superior effect size will not exhibit ceiling (plateau) or floor effects, or clumping along its probability plot. Nor will it show large gaps (which tend to accompany clumping; Chambers, Cleveland, Kleiner & Tukey, 1983; Hintze, 2007). Any of these aberrations reflect lack of disciminability by the effect size index.

Figure 5 depicts the uniform probability distribution for IRD and three comparison effect sizes: PND, Kruskal-Wallis *W* (square root of), and Pearson *R*, based on application to 166 AB contrasts from published datasets. Table 6 elaborates this figure by giving effect size values at 10th, 25th, 50th, 75th, and 90th percentiles. Figure 5 illustrates the superiority of Pearson *R* (clear circle) and Kruskal Wallis *W* (solid circle). Both represent near-diagonal lines, and neither shows clumping or gaps or plateauing. Pearson *R* tops out around .99, and Kruskal-Wallis at values of approximately .85. This fact is important for score interpretation, but is not a deficiency in discriminating among studies. Pearson *R* shows no floor effects, but Kruskal-Wallis shows a little clumping for lowest scores, a small deficiency. The next best distribution is IRD, which ranges from about .10 to 1.0. IRD shows no floor effects, but does show a ceiling effect; it clumps the top 15% of scores together at 1.0, unable to discriminate among them. Least satisfactory is PND, which shows both ceiling and floor effects, clumping the bottom 17% of results together for PND = 0.

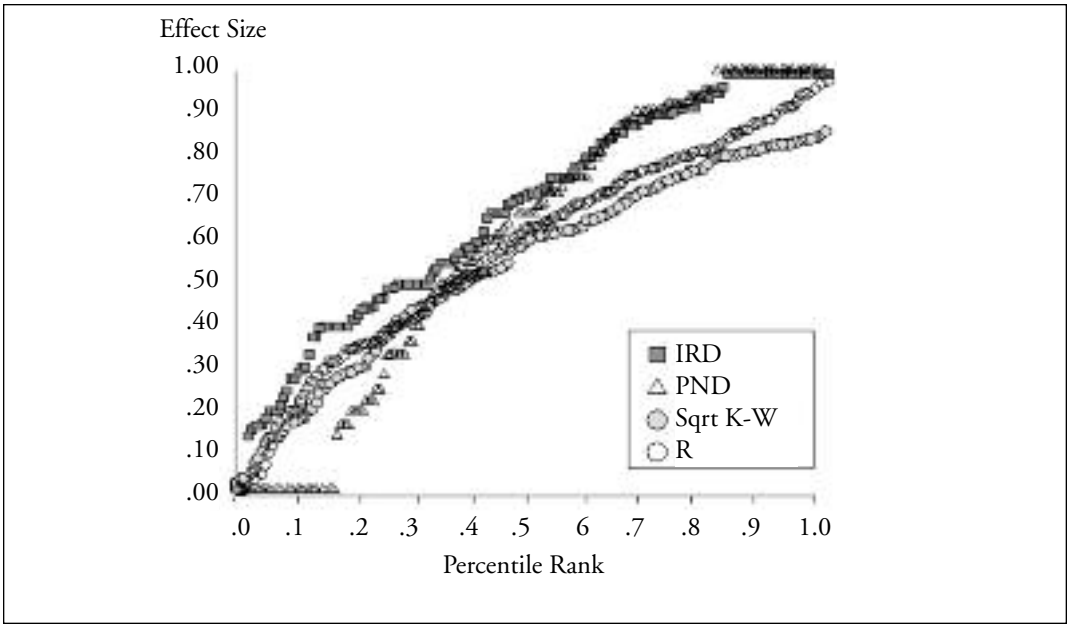
Both IRD and PND identified 29 of the 166 datasets as having no data overlap. However, PND identified 28 additional datasets as having completely overlapping data (PND = 0). In contrast, IRD assigned those 28 datasets IRD values ranging from .17 to .77. Pearson *R* values for those same 28 datasets averaged .38 (range, .04 to .91), more in line with IRD. IRD, Pearson *R* and Kruskal-Wallis *W* were able to detect small-to-large effects in several of the datasets identified as "no effect" by PND.

Table 6 shows IRD effect sizes at key percentiles, beside the other indices. IRD values tended to be about .10 larger than Pearson *R* and Kruskal-Wallis *W* (square root of). The median IRD value for the 166 analyses was .72, with the interquartile range .48 to .90. IRD was .10 larger than Pearson *R* at 90th, 75th, 50th, and 25th percentiles. At the 10th percentile, the difference widens to .17 points.

Reflecting on our example datasets, the Lee et al. (2002) omnibus IRD (for ABAB) of .71 places it at the 50th percentile in Table 5. Lee et al's results therefore appear typical of those found in published single case articles. The Kennedy et al. (1997) multiple baseline omnibus IRD of .83 places it somewhat larger than average—between the 50th and 75th percentiles for the published

FIGURE 5

Uniform Probability Distributions for IRD, PND, Kruskal-Wallis W (Square Root of), and Pearson R, Based on 166 AB Contrasts



studies sampled. In making these comparisons with the sample of 166, we refer only to the magnitude of performance change, not to the strength of the intervention. The extent to which the intervention accounted for client improvements is a matter of research design, not effect size.

DISCUSSION

This article describes a new effect size for single-case research data, IRD, which can be flexibly applied to designs of several phases and series, including multiple baseline designs. IRD, under the name of "risk analysis," has been successfully

TABLE 6

Score Distributions of Seven Effect Size Measures for Single-Case Research, Based on Analyses of 166 Published AB Contrasts

	Percentile				
	10th	*25th*	*50th*	*75th*	*90th*
R Equivalent:					
IRD	.368	.479	.718	.898	.999
Pearson *R*	.200	.375	.626	.795	.888
Sqrt K-W W	.173	.365	.607	.755	.823
PND	.000	.276	.675	.918	.999
R^2 Equivalent:					
IRD^2	.135	.229	.515	.806	.999
Pearson R^2	.040	.140	.392	.632	.789
K-W W	.030	.133	.368	.570	.677

Note. IRD = improvement rate difference; Sqrt K-W W = square root of Kruskal-Wallis W; PND = percent of nonoverlapping data; K-W W = Kruskal-Wallis W.

applied in hundreds of evidence-based medicine group design studies, and is promoted by the international Cochrane Collaborative, and its counterpart in education, the Campbell Collaborative. However, single-case researchers have not yet embraced the robust effect size measures commonly used in the biosciences. Six advantages of IRD are that it (a) is easily calculated by hand; (b) complements single-case research visual analysis; (c) is easily interpreted and explained to lay consumers; (d) has already been established in a respected field; (e) does not require unwarranted data assumptions, as do parametric and even rank order techniques (e.g., Kruskal-Wallis); and (f) has readily obtained confidence intervals. However, more information is required of a new effect size index.

We field-tested IRD with 166 published AB contrasts to answer four additional questions: (a) Can IRD be reliably calculated? (b) How do IRD results relate to better known effect sizes? (c) How well does IRD discriminate among single case datasets? and (d) What IRD effect size magnitudes are typically found? The question of reliable calculation was answered affirmatively by two novice scorers, whose IRD scores on 166 datasets agreed at $R = .96$, with 72% identical scores. This was impressive considering the fact that calculations were done from visual analysis of graphs, some dense with data points. Exact agreement increased to nearly 100% by having each rater recheck the disagreement graphs. The visual analysis step of IRD could be replaced by scrutiny of the data entered into a spreadsheet. For the most dense graphs, scrutinizing a spreadsheet was quicker, but for graphs with fewer data points, it was not. Efficiency of the novice scorers increased markedly quickly. After the first approximately dozen graphs, time was reduced to less than 1 min per graph. Visual analysis, with the assistance of a transparent rule, appears to be sufficiently accurate and efficient for most datasets.

Single-case researchers have not yet embraced the robust effect size measures commonly used in the biosciences.

The second research question was about how IRD results related to better known effect sizes. IRD correlated .86 with the R^2 and Kruskal-Wallis *W* effect sizes, and .83 with the most used index, PND. IRD correlations with R^2 and *W* were over 10 points higher than those achieved by PND. This high-moderate size relationship with the strongest parametric and nonparametric effect sizes lends considerable support to IRD as a new index.

We addressed the third research question, related to IRD's *discriminability* (its ability to differentiate among individual datasets in a reasonably large, typical sample) by uniform probability distributions. IRD showed worse discriminability than R^2 and *W*, but better than PND. IRD showed no floor effects, and discriminated well among the datasets except for the 15% with the largest effect sizes—those with no data overlap between contrasted phases. Those largest effect sizes were all calculated as 1.0 by IRD. These results describe an index sufficiently sensitive for practical use only with designs which show small, medium, and medium-large results.

The final research question was about what IRD effect size magnitudes are typically found. For most of the 166 analyses, IRD values were about .10 larger than Pearson *R* and the square root of Kruskal-Wallis *W*. To many readers, the median IRD value of .72 will seem high compared to a median *R* of .63 (R^2 of .39). However, there are dozens of bona fide effect sizes, and their magnitudes can vary considerably for any given analysis (Cohen, 1988; Parker et al. 2005). Interestingly, some influential statisticians advocate moving from Pearson's R^2 to *R* as a more interpretable effect size, and one whose magnitude better reflects amount of change (Rosenthal et al., 2000). The IRD range of values was close enough to other index values, that an adjustment of interpretation to the new range should not be difficult.

Comparing IRD with PND highlighted some distribution differences and similarities. IRD (along with R^2 and *W*) was able to detect effects (some large) from the AB contrasts PND identified as "no effect." For effect sizes at the 10th percentile, the average PND value was zero, but the average IRD was .37. However, both IRD and PND showed a ceiling effect deficiency, which will exist with any overlap-based index. When there is complete separation of data points between phases, both IRD and PND award the

highest effect size (1.0). Thus, IRD (and PND) should not be used to compare or differentiate studies with large effect sizes.

The two example datasets in the Method section demonstrated a benefit of using IRD with more complex designs: effects from two or more simple phase contrasts can be added together for an omnibus contrast. That is not the case with parametric analyses (R, R^2), where to achieve the same end, variance among noncontrasted phases must be partialed out (Parker & Brossart, 2006). IRD also lends itself to presenting results of multiple phase contrasts together, as in a meta-analysis. In the field of evidence-based medicine, such meta-analytic type displays often feature IRD, termed "risk difference" in medicine. Specialized software such as MetaWin (Rosenberg et al., 2000) and the Cochrane Collaboration's free RevMan (2003) calculate IRD effect sizes and graphically display them from all studies together, in a forest plot (Bijnens, Collette, Ivanov, Hoctin Boes, & Sylvester, 1996). The forest plot shows the results of component studies in a meta-analysis through a visual representation that includes a confidence interval and pooled point estimate to demonstrate significance. The forest plot permits readers, at a glance, to visually analyze 20, 30, or more IRDs—their sizes and reliabilities (via confidence intervals). Forest plots, first used in 1982, are now common in the biosciences, notably in evidence-based medical research (Lewis & Clarke, 2001). There is no reason why such effect size presentation formats cannot also serve single-case research designs.

The detailed IRD demonstration in the Method section shows its flexibility. We calculated IRD for each individual A-versus-B phase shift as well as for all baseline versus all intervention phases together (an omnibus test), and illustrated IRD with a single ABAB series and with a multiple baseline design. IRD can be used to help judge performance change over a series of three or more AB contrasts, supporting guidelines of visual analysis. The IRD procedure relies on visual analysis and hand calculations, with the optional use of statistical software to obtain CIs. Software for calculating IRD and its CI is readily available, including free Internet downloads and a Web-based interactive application. The CIs indicating level of certainty and precision in the obtained IRD may be disappointingly large when calculated for shorter datasets and for smaller IRD values. However, they indicate realistic limitations of short data series with weaker single-case research results.

Ma (2006) recently presented another overlap-based index, the percentage of data points exceeding the median (PEM). In a comparison between PEM and IRD (Parker & Hagan-Burke, 2007), the latter index surpassed PEM in criterion-related validity, including visual analysis ratings about the magnitude of behavior change. Correlations between IRD and visual judgments ranged from .71 to .82 over multiple raters. These are larger than those typically achieved for visual judgments of single-case research (Harbst et al., 1991; Ottenbacher, 1990; Park et al., 1990).

It was beyond the scope of this first IRD study to establish benchmarks for small, medium, and large effects, such as Cohen has done for R^2 in large-N social science research. However, we have calculated IRD on several datasets for which visual analysis ratings were also available. From comparing visual ratings with IRD, we can estimate tentative benchmarks. Very small and questionable effects scored about .50 and below. Moderate-size effects had IRD scores of around .50 to .70. Effects rated as large and very large generally received IRD scores of .70 or .75 and higher.

A limitation of the present study was that we conducted only AB contrasts in the 166 field test contrasts, although similar results are expected for more complex contrasts, as demonstrated in the example data. A caution with IRD is that, like any effect size measure, it does not causally link the intervention and client improvement. Scrutiny of the design is required to make such a causal inference. A second caution relates to positive baseline trend. When Phase A trend is prominent, a calculated effect size cannot fairly represent treatment effectiveness. In those cases, parametric (Allison & Gorman, 1993) or nonparametric (White & Haring, 1980) techniques can control the baseline trend. After applying a trend-compensating formula, IRD can safely be used without modification.

This initial study of IRD (renamed and adapted from the "risk difference" used in medical research) is encouraging. It is a simple, low-effort, low-technology approach very compatible with visual analysis. It has strong interscorer reliability.

It correlated well with the most prestigious parametric and nonparametric effect sizes, and meets APA publication standards of providing confidence intervals. It showed better sensitivity than PND, and was more strongly validated by external measures. This is its first field test, so it needs to be tested further by other researchers with other datasets. However, the present field test with 166 phase contrasts is one of the largest published to date in the field of single-case research.

REFERENCES

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

Altman, D. G. (1999). *Practical statistics for medical research*. Bristol, United Kingdom: Chapman & Hall.

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

Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis, 1*, 91–97.

Bijnens, L., Collette, L., Ivanov, A., Hoctin Boes, G. & Sylvester, R. (1996). Can the forest plot be simplified without losing relevant information in meta-analyses? *Controlled Clinical Trials, 17*, 2S: 124.

Bloom, M., Fischer, J., & Orme, J. G. (1999). *Evaluating practice: Guidelines for accountable professionals* (3rd ed.). Needham Heights, MA: Allyn & Bacon.

Brossart, D. F., Parker, R. I., Olson, E. A., & Mahadevan, L. (2005). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification, 30*, 531–563.

Buchan, I. (2006). StatsDirect (Version 2.5.5) [Computer software]. Cheshire, United Kingdom: Statsdirect.

Busk, P. L., & Marascuilo, L. A. (1992). Statistical analysis in single-case research: Issues, procedures and recommendations, with applications to multiple behaviors. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 159–185). Hillsdale, NJ: Lawrence Erlbaum.

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

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

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

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

Cochrane Collaboration. (2003). Review Manager (RevMan) (Version 4.2) [Computer software]. Copenhagen, Denmark: Author.

Cochrane Collaboration. (2006). *Cochrane handbook for systematic reviews of interventions*. Retrieved May 18, 2006, from http://www.cochrane.org/index_authors_researchers.htm

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

Cooper, J. O., Heron, T. E., & Heward, W. L. (1987). *Applied behavior analysis*. New York: Macmillan.

Durlak, J. A. (2002). Evaluating evidence-based interventions. *School Psychology Quarterly, 17*, 475–482.

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

Giles, T. R. (1990). Bias against behavior therapy in outcome reviews: Who speaks for the patient? *The Behavior Therapist, 13*, 86–90.

Glass, G. V. (1997). Interrupted time series quasi-experiments. In R. M. Jaeger (Ed.), *Complementary methods for research in education.* (2nd ed., pp. 589–608). Washington DC: American Educational Research Association.

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

Harper, R. (1999). Reporting of precision of estimates for diagnostic accuracy: A review. *British Medical Journal, 318*, 1322–1323.

Hintze, J. (2007). NCSS and PASS: Number Cruncher Statistical Systems [Computer software]. Kaysville, UT: NCSS Statistical & Power Analysis Software.

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

Huitema, B. D. (1986). Statistical analysis and single-subject designs. In A. Poling & R. W. Fuqua (Eds.),

Research methods in applied behavior analysis: Issues and advances. (pp. 209–232). New York: Plenum.

Institute of Education Sciences. (2005). *Special education research on individualized education programs.* CFDA number 84.3241. Retrieved May 20, 2006, from http://www.ed.gov/programs/edresearch/applicant.html

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy, 15*, 336–352.

Kennedy, C., Cushing, L. S., & Itkonen, T. (1997). Social contacts and friendship networks of students with severe disabilities. *Journal of Behavioral Education, 7*, 167–189.

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

Kratochwill, T. R., & Brody, G. H. (1978). Single subject designs: A perspective on the controversy over employing statistical inference and implications for research and training in behavior modification. *Behavior Modification, 2,* 291–307.

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

Lee, R., McComas, J. J., & Jawor, J. (2002). The effects of differential and lag reinforcement schedules on varied verbal responding by individuals with autism. *Journal of Applied Behavior Analysis, 35,* 391–402.

Lewis, S., & Clarke, M. (2001). Forest plots: Trying to see the wood and the trees. *British Medical Journal, 322*, 1479–1480.

Linden Software Ltd. (1998). I-Extractor (Version 1.0) [Computer software]. East Yorkshire, United Kingdom: Linden Software.

Lindsley, O. R. (1971). From Skinner to precision teaching: The child knows best. In J. B. Jordan & L. S. Robbins (Eds.), *Let's try doing something else kind of thing* (pp. 1–11). Arlington, VA: Council for Exceptional Children.

Llorca, J. (2002). Computer programs for epidemiologists: PEPI Version 4.0. *Journal of Epidemiology Community Health, 56*, 959–960.

Lovitt, T. C. (1977). *In spite of my resistance . . . . I've learned from children*. Columbus, OH: Charles Merrill.

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

Matyas, T. A., & Greenwood, K. M. (1996). Serial dependency in single-case time series. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 215–243). Mahwah, NJ: Lawrence Erlbaum.

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

Mitchell, M. (2005). dr-ROC (Version 1.1.1) [Computer software]. Glenside, PA: Diagnostic Research Design and Reporting.

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

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

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

Parker, R. I. (2006). Increased reliability for single case research results: Is the bootstrap the answer? *Behavior Therapy, 37*, 248–261.

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

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

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

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

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

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

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

Parsonson, B. S., & Baer, D. M. (1992). The visual analysis of data, and current research into the stimuli controlling it. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis* (pp. 15–40). Hillsdale, NJ: Lawrence Erlbaum.

Petrosino, A., Boruch, R. F., Rounding, C., McDonald, S. & Chalmers, I. (2000). The Campbell Collaboration Social, Psychological, Educational and Criminological Trials Register (C2-SPECTR) to facilitate the preparation and maintenance of systematic reviews of social and educational interventions. *Evaluation and Research in Education, 14*, 206–219.

Rosenberg, M. S., Adams, D. C. & Gurevitch, J. (2000). MetaWin 2.0 (Version 2.0) [Computer software]. Sunderland, MA: Sinauer.

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

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

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

Sackett, D. L., Richardson, W. S., Rosenberg, W., & Haynes, R. B. (Eds.). (1997). *Evidence-based medicine: How to practice and teach EBM*. London: Churchill Livingstone.

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.

Shaver, J. P. (1991). Quantitative reviewing of research. In J. P. Shaver (Ed.), *Handbook of research on social studies teaching and learning* (pp. 83–97). New York: Macmillan.

Sidman, M. (1960). *Tactics of research*. NY: Basic Books.

Skinner, B. F. (1938). *The behavior of organisms*. Englewood Cliffs, NJ: Prentice Hall.

Strube, M. J., Gardner, W., & Hartmann, D. P. (1985). Limitations, liabilities, and obstacles in reviews of the literature: The current status of meta-analysis. *Clinical Psychology Review, 5*, 63–78.

Thompson, B. (2002a). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 3*, 24–31.

Thompson, B. (January, 2002b). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling and Development*. Retrieved July 11, 2005, from http://www.counseling.org/Content/NavigationMenu/PUBLICATIONS/JOURNALS/JOURNALS.htm

Thompson, B. (2006). *Effect sizes, confidence intervals, and especially confidence intervals for effect sizes*. American Educational Research Association annual meeting professional development course. Retrieved May 22, 2006, from http://www.aera.net/annualmeeting/?id=294

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

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.

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

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

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

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

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

Wolf, F. M. (2000). Lessons to be learned from evidence-based medicine: Practice and promise of evidence-based medicine and evidence-based education. *Medical Teacher, 22*, 251–259.

ABOUT THE AUTHORS

RICHARD I. PARKER (CEC TX Federation), Associate Professor; and **KIMBERLY J. VANNEST** (CEC TX Federation), Assistant Professor, Educational Psychology, Texas A&M University, College Station. **LEANNE BROWN** (CEC TX Federation), Special Education Teacher, Cox Elementary School, Cedar Park, Texas.

Address correspondence to Richard Parker, 704 Harrington Tower, Texas A&M University, College Station, TX 77843-4225 (e-mail: rparker@tamu.edu).

Manuscript received June 13, 2006; accepted April 21, 2008.