

*VISUAL AIDS AND STRUCTURED CRITERIA FOR
IMPROVING VISUAL INSPECTION AND INTERPRETATION OF
SINGLE-CASE DESIGNS*

WAYNE W. FISHER

MARCUS AND KENNEDY KRIEGER INSTITUTES AND
JOHNS HOPKINS UNIVERSITY SCHOOL OF MEDICINE

MICHAEL E. KELLEY

MARCUS INSTITUTE AND
LOUISIANA STATE UNIVERSITY

AND

JOANNA E. LOMAS

MARCUS INSTITUTE

Because behavior analysis is a data-driven process, a critical skill for behavior analysts is accurate visual inspection and interpretation of single-case data. Study 1 was a basic study in which we increased the accuracy of visual inspection methods for A-B designs through two refinements of the split-middle (SM) method called the dual-criteria (DC) and conservative dual-criteria (CDC) methods. The accuracy of these visual inspection methods was compared with one another and with two statistical methods (Allison & Gorman, 1993; Gottman, 1981) using a computer-simulated Monte Carlo study. Results indicated that the DC and CDC methods controlled Type I error rates much better than the SM method and had considerably higher power (to detect real treatment effects) than the two statistical methods. In Study 2, brief verbal and written instructions with modeling were used to train 5 staff members to use the DC method, and in Study 3, these training methods were incorporated into a slide presentation and were used to rapidly (i.e., 15 min) train a large group of individuals ($N = 87$). Interpretation accuracy increased from a baseline mean of 55% to a treatment mean of 94% in Study 2 and from a baseline mean of 71% to a treatment mean of 95% in Study 3. Thus, Study 1 answered basic questions about the accuracy of several methods of interpreting A-B designs; Study 2 showed how that information could be used to increase the accuracy of human visual inspectors; and Study 3 showed how the training procedures from Study 2 could be modified into a format that would facilitate rapid training of large groups of individuals to interpret single-case designs.

DESCRIPTORS: assessment, behavior analysis, data analysis, interrater agreement, visual inspection

One area of behavioral research that continues to be a critical component of applied behavior analysis is staff training (Page, Iwata, & Reid, 1982; Reid & Parsons, 1995), be-

cause accurate implementation of behavioral principles and procedures is critical to their effectiveness. However, indirect teaching procedures (ones in which the participant is not required to emit the target response; e.g., written instructions, lectures) are often the primary methods used to train new staff members in behavior analysis procedures, even though direct methods (ones in which the participant is required to emit the target response; e.g., direct instruction, modeling,

We thank Amanda J. Oberdorff for her assistance with this project. This investigation was supported in part by Grant 5 R01 HD37837-03 from the Department of Health and Human Services, the National Institute of Child Health and Human Development.

Requests for reprints should be sent to Wayne W. Fisher, Marcus Behavior Center, 1920 Briarcliff Road, Atlanta, Georgia 30329.

behavioral rehearsal, feedback) often result in better training outcomes (see Watson & Kramer, 1995). Indirect methods are probably used more often, because direct methods tend to be more costly and time consuming.

The current investigation is part of an ongoing line of research designed to identify critical instructional components for training new staff members in the implementation of behavior-analytic procedures, with the goal of approximating the efficiency of indirect instructional methods while retaining the effectiveness of more direct methods. In the current study, we focused on training staff members to interpret behavioral data presented in A-B designs. We selected this skill because applied behavior analysis is a data-driven process, and accurate visual inspection and interpretation of single-case data are essential to the successful practice of applied behavior analysis. We do not endorse reliance on A-B designs, and visual inspection should always consider the context in which the clinical or scientific question is posed. We chose interpretation of A-B designs because determining whether a reliable change in behavior has occurred between baseline and treatment is a prerequisite to determining whether that effect is reversed during a treatment withdrawal or replicated during a treatment reinstatement.

Although behavior analysts have suggested that visual inspection of single-case data is generally reliable and conservative (Baer, 1977; Michael, 1974; Parsonson & Baer, 1986), findings from empirical studies on this subject have suggested otherwise (for a review, see Franklin, Gorman, Beasley, & Allison, 1996). For example, DeProspero and Cohen (1979) found an interrater-agreement coefficient of just .61 (Pearson correlation) among individuals who reviewed articles for publication in behavioral journals. Other studies have found similar or even lower levels of interrater agreement among less experienced visual inspectors (e.g., Boy-

kin & Nelson, 1981; Harbst, Ottenbacher, & Harris, 1991; Ottenbacher, 1990).

Several investigations have focused on improving the reliability of visual inspection methods by training judges to apply structured criteria (Hagopian *et al.*, 1997) or through the provision of visual aids (Bailey, 1984; Rojahn & Schulze, 1985). The structured criteria developed by Hagopian *et al.* improved both the reliability and the validity of interpretations made by graduate students in an internship training program (when compared with the visual interpretations of an expert panel). However, the Hagopian *et al.* criteria are not well suited for the purposes of the current investigation because these criteria were specifically developed for multielement designs and the training procedures are somewhat cumbersome and time consuming.

One simple and efficient method of increasing the reliability and validity of visual inspection involves the provision of visual aids in the form of trend lines (or lines of progression). This method has produced modest increases in the reliability of visual inspection (Bailey, 1984) and in agreement levels between visual inspection and statistical analyses (Rojahn & Schulze, 1985). In the Rojahn and Schulze study, one linear regression line (least squares estimate using the slope and intercept) was generated from the baseline data and was superimposed on the baseline phase; a separate regression line was generated from the treatment data and was superimposed on the treatment phase. In the Bailey investigation, a trend line was superimposed on the treatment phase that was based on the baseline data path to help determine whether the treatment data appeared to be a continuation of the data path that followed the trend established in baseline. The lines of progression in the Bailey study were generated using the split-middle (SM) method (Kazdin, 1982; Parsonson & Baer, 1986; White, 1974), which is a quick method of estimating a least squares linear regression line.

Kazdin (1982) suggested that one could (a) apply the SM method, (b) count the number of treatment points that fell above (or below) the trend line superimposed on the treatment phase, and then (c) apply the binomial formula to calculate the probability of that number (of data points falling above the line) occurring by chance. It is likely that applying the SM method with the binomial formula as a decision aid, as Kazdin recommended, would have greatly improved the reliability of visual inspection, whereas the SM line alone produced only modest gains (Bailey, 1984). However, Crosbie (1987) found that the accuracy of the binomial test is decreased markedly by the presence of serial dependence in the data series. Serial dependence in behavioral data is said to occur when the level of behavior at one point in time (e.g., increased sleep on the weekend) is either influenced by or correlated with the level of behavior at a previous point in time (e.g., reduced sleep during the work week). Because serial dependence occurs in single-case data series (although authors disagree as to the extent to which it occurs; for a review, see Matyas & Greenwood, 1996), the SM method, used as Kazdin recommended, would probably increase the reliability but perhaps not the validity of visual inspection methods. In addition, applying the binomial test in combination with the SM method may not be appropriate when the baseline phase shows an apparent trend (Kazdin, 1982).

In addition to problems with reliability, some investigators have suggested that the validity of visual inspection is poor, because when visual and statistical procedures have been directly compared, agreement between the two methods has been low (Jones, Weinrott, & Vaught, 1978; Park, Marascuilo, & Gaylord-Ross, 1990; Rojahn & Schulze, 1985). However, studies comparing visual inspection and statistical methods for single-case experiments have had a number of de-

sign issues that may have limited the generality of the findings (Matyas & Greenwood, 1990). For example, Jones et al. specifically selected A-B phases from figures published in the *Journal of Applied Behavior Analysis (JABA)* with relatively small or no treatment effects and ones that appeared to show evidence of serial dependence (among other selection criteria). These selection biases probably deflated agreement levels between the visual and statistical methods. Conversely, Park et al. randomly selected A-B phases from figures published in *JABA*, but included only ones with at least 25 data points between the two phases, which resulted in a bias toward A-B phases with no treatment effects and may have inflated the overall agreement between the two approaches. Interestingly, when disagreements occurred between the visual and statistical interpretations, Jones et al. found that visual inspection tended to be more conservative whereas Park et al. did not.

Perhaps the most important limitation of prior studies that have compared the results of visual inspection and statistical analyses on the same data sets is that when the two methods produce discordant interpretations, it was not possible to determine which interpretation was correct (Matyas & Greenwood, 1990). Matyas and Greenwood attempted to overcome this limitation by having visual inspectors interpret graphs for which the "correct" interpretation was known beforehand, rather than using a particular statistical procedure as a "gold standard." They had a group of 37 graduate students in a course on single-case designs interpret 27 A-B graphs created using a first-order autoregressive model: $Y_i = aY_{i-1} + B + D + E$, where Y_i was the dependent variable at time i , Y_{i-1} was the dependent variable at time $i - 1$, a was the autocorrelation coefficient (which controlled whether or not the data were serially dependent), B was the baseline mean, D was the intervention effect,

and E was error. This allowed the authors to program whether there was a treatment effect (i.e., $D = 0, 5$, or 10) and whether the data were serially dependent (i.e., $a = 0, 0.3$, or 0.6). They then used these data to estimate the rates of Type I and Type II errors. Type I errors are ones in which a conclusion is made that a treatment (or other independent variable) produced a real change in behavior, when in fact, the change was due to chance or other variables. Type II errors are ones in which a conclusion is made that a treatment did not produce a real change in behavior, when in fact, it did.

Results of the Matyas and Greenwood (1990) investigation indicated that a high percentage of the judges (16% to 84%) made Type I errors when the amount of autocorrelation was greater than zero and the amount of random error was three standard deviations or greater, whereas Type II errors tended to occur less often (0% to 22% of judges across graphs).

Although the Matyas and Greenwood (1990) investigation represented a significant methodological advancement over prior studies designed to determine the accuracy of visual inspection methods, it was limited in that only one graph was created (i.e., sampled) from each set of autoregressive parameters. When similar autoregressive models have been used to determine rates of Type I and Type II errors for statistical procedures, each combination of parameters (e.g., $a = 0$, $B = 10$, $D = 5$) has generally been used to create thousands of data samples and the statistical procedure was applied to each sample using Monte Carlo methods (e.g., Ferron & Ware, 1995). Using a large number of samples for each set of parameters insures that the Type I and Type II error rates are accurate and are not the result of sampling error. However, it would not be practical to have humans visually inspect thousands of graphs to ascertain their rates of Type I and Type II errors. Nevertheless, by

generating just one sample graph from each set of parameters, Matyas and Greenwood left open the possibility that some of the graphs were not representative of the model parameters used to create those graphs.

To examine the possibility that Matyas and Greenwood's (1990) results were influenced by sampling error, we applied a statistical test to the data set that proved most problematic for the visual inspectors in that study ($a = 0.3$, $S = 5$, $D = 0$; see Figure 1 from Matyas & Greenwood, p. 344). The statistical test was a general linear model (GLM) using the same autoregressive model as Matyas and Greenwood used to generate the graph, and the GLM was designed to answer the same question asked of the visual inspectors. That is, was there a reliable treatment effect, or in statistical terms, was the null hypothesis that $D = 0$ false? When we applied the GLM to this sample ($a = 0.3$, $S = 5$, $D = 0$), the resultant F value was 7.7 ($p = .01$ with 1 and 16 df).

Our statistical analysis suggested that this particular sample graph was not representative of the parameters that were used to create it. Moreover, when viewed relative to the results of this statistical analysis, the results obtained by the judges in the Matyas and Greenwood (1990) investigation for this sample graph do not seem nearly as troublesome as they otherwise would. That is, Matyas and Greenwood had 37 graduate students visually inspect a single graph generated with the population parameters $a = 0.3$, $S = 5$, $D = 0$, and 31 of the students incorrectly concluded that there was a treatment effect (i.e., they concluded that D was not equal to 0). Based on this result, Matyas and Greenwood concluded that the Type I error rate was 84% for these visual inspectors. If we applied the GLM statistical test 37 times to this same data set, it would have produced an incorrect interpretation each time; however, it would not be reasonable to conclude from these results that the Type I

error rate for the statistical test was 100%. Similarly, it is not reasonable to conclude that the Type I error rate for visual inspection is 84% because 31 of 37 graduate students incorrectly concluded that there was a reliable treatment effect for this one, unrepresentative, graph. These findings suggest that Type I and Type II error rates for visual inspection methods should be evaluated using Monte Carlo procedures that are equivalent to those used to evaluate statistical methods. One way that this might be practically accomplished would be to use a visual inspection guide, like the SM method, for which it would be possible to simulate the method with a computer, which could rapidly and repeatedly inspect thousands of data sets.

From this selected review of the literatures on staff training and visual and statistical interpretation of single-case data, we have generated the following assumptions. First, further refinement of training methods, structured criteria, and visual aids is needed to improve the reliability and validity of visual inspection methods. Second, Monte Carlo methods should be used to determine the effects these refinements have on the rates of Type I and Type II errors for visual inspection of single-case data. Third, the rates of Type I and Type II errors for the refined methods should be compared to existing visual inspection procedures (e.g., the SM method) as well as statistical methods that have been proposed as alternatives or adjuncts to visual inspection (e.g., least squares GLM, Gorman & Allison, 1996; interrupted time series [ITSE], Gottman, 1981). Finally, for the refinements of visual inspection procedures to be integrated into routine staff training programs (e.g., during job orientation), the training method, criteria, and aids need to be simple and efficient in terms of training time and costs. Studies 1, 2, and 3 were designed with consideration of these assumptions.

STUDY 1: DEVELOPMENT AND VALIDATION OF THE DUAL CRITERIA METHOD

METHOD

Development of the Refined Visual Inspection Methods

We developed a refinement of the SM method, which we call the dual-criteria (DC) method, and a more conservative refinement of the DC method, called the conservative dual-criteria (CDC) method. The refinement process began by first constructing a program in ExcelTM that generated (a) data sets (and accompanying graphs) using a first-order autoregressive model (see the definition of the model below) and (b) a regression line from the baseline data that was superimposed on the treatment phase (i.e., the SM method). We created graphs with and without programmed treatment effects to examine patterns of Type I and Type II errors made by the SM method. Through visual inspection, we noticed that the SM method tended to make Type I errors primarily when the observed data path in baseline was on either an upward or a downward trend, but the observed slope was not programmed by the autoregressive model (i.e., the observed slope in baseline was due to sampling error). When an erroneous downward trend appeared in the baseline but not in the treatment phase, the SM method often incorrectly concluded that a treatment effect had occurred. To compensate for this deficiency in the SM method, we added a second criterion line that was generated from the baseline mean and superimposed on the treatment phase. This resulted in two criteria: (a) A prespecified number of treatment data points had to fall above (or below) the trend line based on the binomial test (as with the SM method), and (b) the same number of data points also had to fall above (or below) the mean line (see the top and bottom panels of Figure 1 for a sample data set with and without these criterion

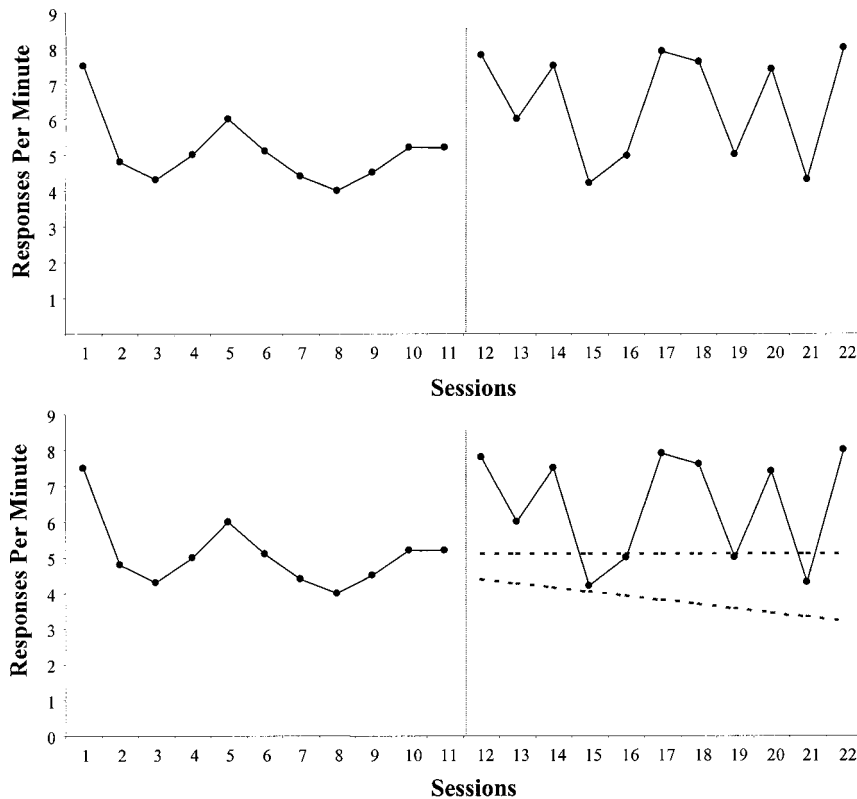


Figure 1. The top panel shows a computer-generated A-B graph without visual aids; the bottom panel shows the same graph with the dual-criteria (DC) visual aids.

lines). The number of data points that had to fall above both criterion lines was based on the binomial equation, the same as for the SM method. We also created a more conservative version of the DC method (the CDC) by raising the height of the two criterion lines by 0.25 standard deviations (calculated from the baseline data). This more conservative version was created after we tested the DC method at several different levels of autocorrelation and found that its rate of Type I errors was unacceptably high. We then probed raising the height of the two criterion lines by several different values and judged that 0.25 standard deviations represented a reasonable compromise between Type I and Type II errors.

Monte Carlo Validation of the DC Method

Four pairs of Excel[™] worksheets were created to conduct the Monte Carlo study.

Each worksheet was created independently of its pair to check the accuracy of the Monte Carlo data generated by the worksheets. Each worksheet was tested against its pair using Monte Carlo simulation methods with 30,000 repetitions until the two worksheets produced nearly identical results (i.e., identical values when rounded to a two-digit number; e.g., 0.054 and 0.048).

One pair of worksheets was designed to determine the error rates for graphs with 10 data points (five per phase), and a second pair was designed for graphs with 20 points (10 per phase). The third and fourth pairs of worksheets were created for graphs with 10 and 20 data points, respectively, after the DC method was adjusted to be more conservative (the CDC method). The lengths of the graphs were chosen because they are typical of ones seen in

articles published in *JABA* (Huitema, 1985).

Each worksheet created a data set using the autoregressive model $Y_i = B + D + (1 - a)E_i + aE_{i-1}$, where Y_i was the dependent variable at time i , B was the baseline mean (which was always set at 10), D was the intervention effect, E_i was random error at time i , a was the autocorrelation parameter, and E_{i-1} was the error term at time $i - 1$.

To test error rates at different levels of autocorrelation, the autocorrelation parameter (a) was alternately set at 0 (which produced uncorrelated error terms), 0.1 (which produced a first-order autocorrelation of 0.11 when tested with an array of 10,000 data points), 0.3 (which produced a first-order autocorrelation of 0.37), and 0.5 (which produced a first-order autocorrelation of 0.5). The random error term (E_i) was generated with the random number function from the add-in program called Resampling Stats[™] for Excel[™], which produced a normally distributed set of random numbers with a mean of 0.0 and a standard deviation of 1.0. In the autoregressive model, E_i was multiplied by $(1 - a)$ so that the two error terms $[(1 - a)E_i + aE_{i-1}]$ produced a sum with a mean of 0.0 and standard deviation of 1.0, which allowed us to vary effect sizes [D divided by the standard deviation of the sum of $(1 - a)E_i + aE_{i-1}$] linearly by adding a constant (0.5) to the effect-size parameter (D). Accordingly, the parameter for the intervention effect (D) was set at 0.0 to determine the rates of Type I errors and alternately was set at 0.5, 1.0, 1.5, 2.0, 2.5, and 3.0 to determine the power of (and the Type II error rates for) the three visual inspection methods (SM, DC, and CDC) and the two statistical analysis methods (GLM and ITSE) described below.

As indicated above, we systematically manipulated two of the parameters in the autoregressive model (the intervention effect, D , and the level of autocorrelation, a). For each combination of these two parameters

(e.g., $D = 0$, $a = 0.3$), the repeat-and-score function of the Resampling Stats[™] program was used to generate 30,000 sample data sets from the autoregressive model. Each time a sample data set was generated, the sample was evaluated using the three visual inspection methods (SM, DC and CDC) and the two statistical analysis methods (GLM and ITSE) described below.

The programs calculated the Type I error rates for the interpretative procedures (when $D = 0$) by counting the number of times each procedure incorrectly concluded that a reliable treatment effect was present in the sample data sets and then dividing that number by 30,000 to get a proportion. The programs calculated the power ($1 -$ the proportion of Type II errors) of each of the interpretative procedures (when $D > 0$) by counting the number of times each procedure correctly concluded that a reliable treatment effect was present in the sample data sets and then dividing that number by 30,000 to get a proportion.

It should be noted that for the three visual inspection methods and two statistical analysis methods, only positive values for the treatment-effects parameter (D) were programmed. Therefore, a one-tailed test was evaluated for each procedure.

GLM. The GLM method used in the current investigation was similar to those recommended to test effect sizes for single-case experiments by Center, Skiba, and Casey (1985–1986) and Allison and Gorman (1993). The full model was $Y_i = b_0I_B + b_1I_T + b_3S_B + b_4S_T + E_i$, where Y_i was the dependent variable at time i , b_0I_B was the least squares predictor term for the intercept for the baseline phase, b_1I_T was the intercept for the treatment phase, b_3S_B was the slope of the baseline phase, b_4S_T was the slope of the treatment phase, and E_i was random error. This full model was tested against a restricted model that forced the baseline and treatment phases to have the same intercept and

slope ($Y_i = b_0I + b_1S + E_i$). The sum-of-squares error terms from these two models were then compared using the formula $F = (SSE_r - SSE_f/df_1)/(SSE_r/df_2)$, where SSE_r was error sum of squares for the restricted model, SSE_f was the error sum of squares for the full model, df_1 was the number of predictor terms in the full model minus the number of predictor terms in the restricted model ($df_1 = 4 - 2 = 2$), and df_2 was the number of data points (i.e., either 10 or 20) minus the number of predictor terms in the full model ($df_2 = 6$ for the 10-point graphs and 16 for the 20-point graphs).

ITSE. The ITSE method used in the current investigation was equivalent to the omnibus F test described by Gottman (1981). The full model was $Y_i = b_0I_B + b_1I_T + b_3S_B + b_4S_T + b_5Y_{i-1} + E_i$. This full model was identical to the one described above for the GLM procedure except that it also included a predictor term to estimate the level of first-order autocorrelation (b_5Y_{i-1}), or the extent to which behavior at time i was related to behavior at time $i - 1$. This full model was tested against a restricted model that forced the baseline and treatment phases to have the same intercept and slope but retained the term for the first-order autocorrelation ($Y_i = b_0I + b_1S + b_2Y_{i-1} + E_i$). The sum-of-squares error terms for the full and restricted ITSE models were then compared using the formula for the F test described above for the GLM except that a different value was used for df_2 to account for the inclusion of an autoregressive parameter in the model (b_5Y_{i-1}). For the 10-point graphs, df_2 equaled 4, and for the 20-point graphs, df_2 equaled 14 (because autoregressive parameters use up 2 degrees of freedom rather than 1).

SM method. Because we were conducting a Monte Carlo study in which 600,000 data sets were interpreted (i.e., 20 autoregressive models $\times$ 30,000 repetitions), it was not practical to have humans actually visually inspect each graph. Therefore, the worksheets

were programmed to perform the same tasks required of a human visual inspector when using the SM method as described by Kazdin (1982). The worksheets were programmed to calculate the SM trend line based on the baseline data points using the least squares formula, and it also counted the number of treatment data points that fell above that line. If all five data points in the treatment phase of a 10-point graph fell above the SM criterion line, the program scored the graph as having a positive treatment effect (just as visual inspectors would do when using the SM method); otherwise, it was scored as not having a treatment effect. For a 20-point graph, at least eight of the treatment data points had to fall above the SM criterion line for a positive treatment effect to be scored.

DC method. The worksheets were also programmed to calculate the mean of the baseline data points, to create a second criterion line based on that baseline mean, and to count the number of treatment data points that fell above that line. If all five data points in the treatment phase of a 10-point graph fell above both the SM criterion line and the mean criterion line, the program scored the graph as having a positive treatment effect for the DC method; otherwise, it was scored as not having a treatment effect. For a 20-point graph, at least eight of the treatment data points had to fall above both the SM criterion line and the mean criterion line for a positive treatment effect to be scored.

CDC method. We programmed two worksheets (one for 10-point and one for 20-point graphs) to create this more conservative (i.e., fewer Type I errors) version of the DC method by raising the height of the SM criterion line and the height of the mean criterion line by 0.25 standard deviations (calculated from the baseline data points). If all five data points in the treatment phase of a 10-point graph fell above both the adjusted SM criterion line and the adjusted mean criterion line, the program scored the graph as having

a positive treatment effect for the CDC method; otherwise, it was scored as not having a treatment effect. For a 20-point graph, at least eight of the treatment data points had to fall above both the adjusted SM criterion line and the adjusted mean criterion line for a positive treatment effect to be scored.

RESULTS AND DISCUSSION

The top left panel of Figure 2 shows the proportion of Type I errors for each of the five interpretive procedures when there was no programmed treatment effect ($D = 0$), the level of programmed autocorrelation (a) varied between 0 and 0.5, and there were 20 data points (10 per phase). The SM method produced high error rates across all levels of autocorrelation. The DC and GLM methods produced tolerable error rates when there was no autocorrelation, but error rates increased to unacceptable levels as the level of autocorrelation increased. The ITSE procedure produced tolerable error rates across all levels of autocorrelation. The CDC method was negatively affected by autocorrelation, but even so, it produced lower rates of Type I errors than the other four interpretive procedures at all levels of autocorrelation.

The top right panel of Figure 2 shows the proportion of Type I errors for each of the five interpretive procedures when there was no programmed treatment effect ($D = 0$), the level of programmed autocorrelation (a) varied between 0 and 0.5, and there were 10 data points (five per phase). The SM method again produced high error rates across all levels of autocorrelation. The GLM procedure again produced tolerable error rates when there was no autocorrelation, but error rates increased to unacceptable levels as the level of autocorrelation increased. The DC method was more conservative with the 10-point data sets than with the 20-point data sets (presumably because all five treatment data points had to fall above both criterion lines), producing error rates only slightly above tolerable levels (0.07)

when the level of autocorrelation reached 0.5. The ITSE program produced tolerable error rates with low levels of autocorrelation, but error rates were elevated somewhat (0.08) when the level of autocorrelation was 0.5. The CDC method again produced lower rates of Type I errors than the other four interpretive procedures and was only slightly affected by autocorrelation.

The bottom left panel of Figure 2 shows the proportion of correctly detected treatment effects for each of the interpretive procedures applied to graphs with 20 data points when the treatment effect size (D) varied between 0.5 and 3.0 with no autocorrelation ($a = 0$). Not surprisingly, the SM method, which had the highest Type I error rates, showed the highest power levels, especially when effect sizes were small (e.g., $D = 0.5$). However, the other two visual inspection methods, DC and CDC, showed sharp increases in power levels as the effect size increased such that all three visual inspection methods (SM, DC, and CDC) reached conventionally desirable power levels (0.8 or above) when the effect size was 2.0. By contrast, the two statistical methods, GLM and ITSE, showed low power levels when effect sizes were small, as expected, but surprisingly, did not reach desirable levels until the effect size reached 3.0.

As can be seen in the bottom right panel of Figure 2, the difference between the power levels of the visual inspection and statistical methods were even more pronounced when these interpretive procedures were applied to graphs with 10 data points. It should be noted that all of the interpretive procedures showed lower power for graphs with 10 data points relative to the graphs with 20 data points. Two of the visual inspection methods (SM and CD) produced desirable power levels (above 0.80) when the effect size reached 3.0, and the third (CDC) was only slightly below that level (i.e., 0.79). By contrast, the two statistical methods

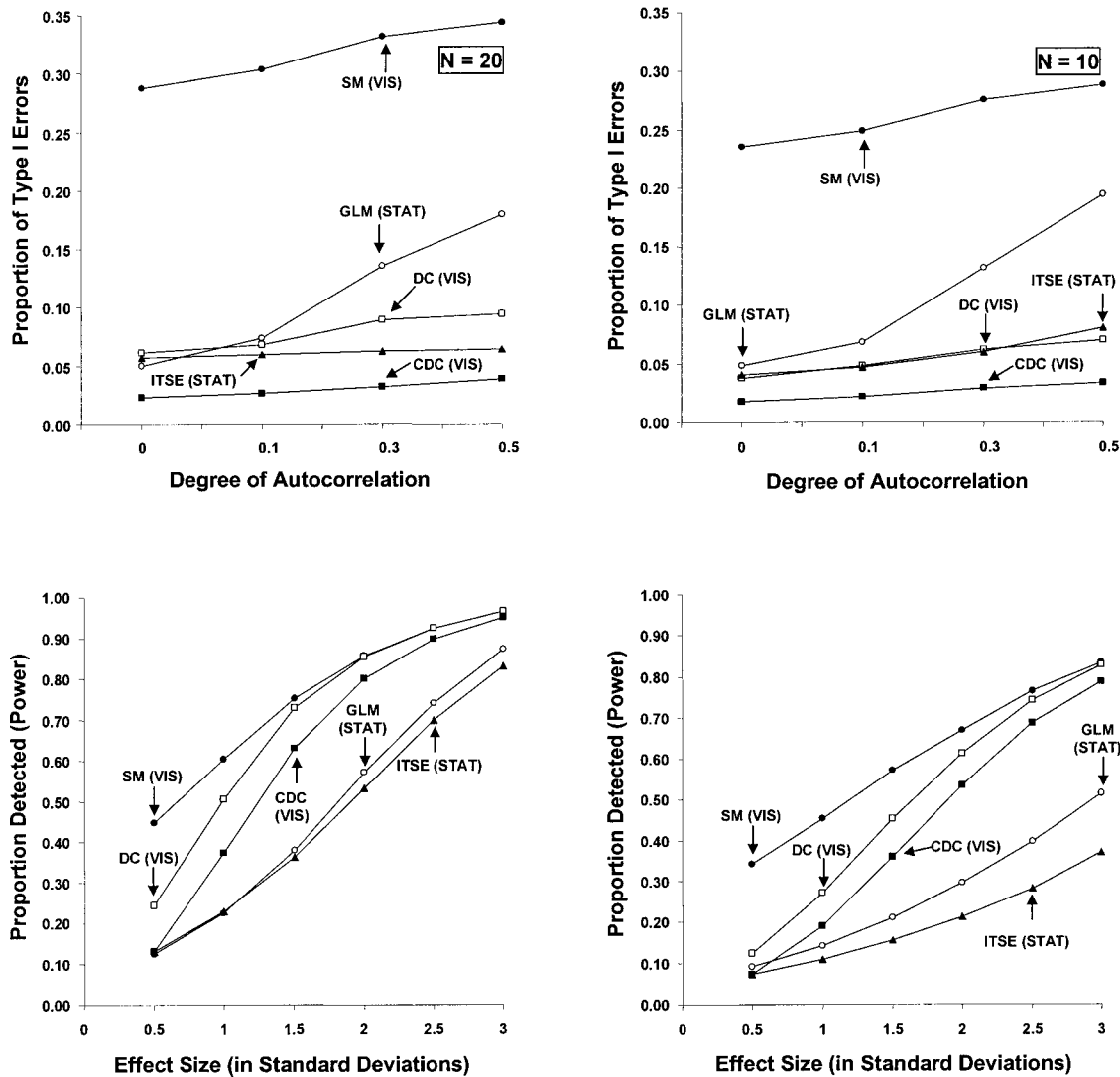


Figure 2. The top panels show the rates of Type I errors at various levels of autocorrelation for the split-middle (SM), dual-criteria (DC), and conservative dual-criteria (CDC) visual inspection methods (VIS), and for the general linear model (GLM), and interrupted time series (ITSE) statistical methods (STAT) of interpreting single-case designs for graphs with 20 points and 10 points. The bottom panels show the power levels ($1 -$ the proportion of Type II errors) for these five interpretive methods at various effect sizes for graphs with 20 points and 10 points.

(GLM and ITSE) obtained power levels of only 0.51 and 0.37, respectively, when the effect size was 3.0.

We identified limitations of the SM method for guiding visual inspection by creating computer-generated graphs with known autoregressive parameters and observing patterns of Type I errors. We used this infor-

mation to develop two refinements of the SM method called the DC and CDC methods. We then compared these three visual inspection methods with each other and with two statistical methods commonly recommended for use with single-case designs, GLM (Gorman & Allison, 1996) and ITSE (Gottman, 1981). Results of the Monte Car-

lo investigation showed that Type I error rates were (a) universally high for the SM method; (b) elevated to unacceptable levels for the GLM procedure for both 10- and 20-point graphs when autocorrelation was 0.3 or higher; (c) elevated to unacceptable levels for the DC method for 20-point graphs when autocorrelation was 0.3 or higher; (d) elevated to unacceptable levels for the ITSE program only for 10-point graphs and only when autocorrelation was 0.5; and (e) at or below tolerable levels for the CDC method for all data-set lengths and levels of autocorrelation tested. Results of the Monte Carlo study also indicated that the three visual inspection methods showed higher power levels than the two statistical methods. Taken together, these results indicate that the CDC method was the only interpretive procedure that both (a) controlled Type I error rates for short data series with and without autocorrelation and (b) maintained reasonable power levels.

STUDY 2: TRAINING VISUAL INSPECTORS IN THE DC METHOD

The purpose of Study 2 was to determine whether staff members could be trained to apply the DC method and whether that would improve the accuracy of their interpretations of A-B single-case designs. The results of Study 1 indicated that the CDC method was the preferred method (among those tested) of interpreting A-B single-case designs. However, in Studies 2 and 3, we trained the participants in the DC method rather than the CDC method. We did this because the time frames for Studies 1, 2, and 3 overlapped and the results of Study 1 were not available at the outset of Studies 2 and 3. Nevertheless, the target responses required of the visual inspector are identical for the DC and CDC methods, so that results of

Studies 2 and 3 should be applicable to either method.

METHOD

Materials

A modified version of the Excel[™] worksheets described and used in Study 1 was used to generate and print graphs. The autoregressive model was the same as used in Study 1; however, the program was modified to produce graphs of several different lengths (i.e., 8, 10, 12, 14, 16, 20, 30, and 40 data points). First, graphs were eliminated if interpretation with the DC method resulted in an error (e.g., the DC method concluded that there was a treatment effect when the graph was drawn from an autoregressive model with $D = 0$). These graphs were eliminated because the purpose was to train the participants to implement the DC method accurately. In addition, graphs with highly obvious treatment effects were eliminated (ones with no overlapping data points, little variance, no observed slope in either phase, and clear differences between means of the phases). Then, the remaining printed graphs were categorized into four groups. Group A consisted of graphs that met both of the DC criteria (described in Study 1). Group B was comprised of graphs that met only one of the DC criteria. Group C included only graphs that fell just one data point short of meeting one of the DC criteria. All other graphs were placed in Group D. Next, 12 packets (20 graphs each) were constructed so that each packet included 10 graphs from Group A, 4 from Group B, 3 from Group C, and 3 from Group D. Packets were constructed in this manner so that (a) each packet contained 10 graphs that showed a treatment effect and 10 that did not, and (b) the difficulty level of each packet was approximately equal to the difficulty level of the other packets.

Participants and Setting

Participants were 5 behavior therapists with baccalaureate degrees employed at a fa-

cility that specialized in the assessment and treatment of severe behavior problems. All participants had at least 4 months of daily exposure to visual inspection and interpretation of single-case designs. All baseline sessions, training, and treatment sessions were conducted in a typically furnished office. Participants were seated at a desk and were provided with a pen or pencil, an answer sheet, and a packet of 20 graphs. Each graph was printed on a piece of paper that measured 21.5 cm by 28 cm.

Response Measurement and Reliability

Participants recorded their interpretations on a preprinted answer sheet by circling the word "yes" if they judged that there was a reliable treatment effect and "no" if they judged otherwise. An experimenter scored the accuracy of participant responses, and a second experimenter independently scored 33% of the participant answer sheets to ascertain the reliability of the scoring method, which was 100%.

Procedure

Prior to conducting a session, the experimenter presented a 20-graph packet to the participant. The order of packet presentation was determined randomly for each participant prior to the study. The participants had unlimited time to judge each of the 20 graphs (but sessions generally lasted about 5 to 15 min). Prior to the start of the session, the experimenter explained how to mark answers on the data sheet and told the participant to mark "yes" whenever he or she judged a treatment effect to be reliable, even if it was a small but reliable effect. The participants were told that being able to detect small but reliable treatment effects is sometimes important, and they were given the example that small decreases in mean blood pressure can produce clinically important health benefits.

Baseline. The experimenter gave a packet of 20 graphs that were similar in form to the

one depicted in the top panel of Figure 1 and also presented the above instructions. When the participant was finished, the experimenter removed the 20 graphs and either initiated another session or terminated the sessions for the day. No more than three sessions were conducted on a given day.

Training. Training was conducted between the baseline and treatment phases. Training consisted of the following components. First, the experimenter presented a graph that was similar in form to the one depicted in the bottom panel of Figure 1. Like that graph, this first training graph (and all subsequent graphs used in training or treatment sessions) included the DC criteria lines (one generated using the baseline mean and the other using the least squares trend line based on the baseline intercept and slope). That is, in addition to the treatment data, the treatment phase of each graph contained two dashed lines, one of which indicated the mean of the baseline data and the other the least squares trend line of the baseline data. The participants were given a modified data sheet that included (a) a table, similar to Table 1, showing the number of data points that needed to be above both criterion lines for graphs with various phase lengths and (b) the rules for using the criterion lines. The rules were written on the back of the data sheet. The rules instructed the participants (a) to count and write down the number of treatment sessions, (b) to count the number of data points in the treatment phase that were above both of the dashed criterion lines, and (c) to look up (in a table similar to Table 1) the number of data points that needed to be above both criterion lines to conclude that a treatment effect was present in the data set.

The participants were then shown two sample graphs. In one graph, the treatment data met both criteria for a difference between baseline and treatment. In the other graph, the treatment data did not meet both criteria for a difference between baseline and

Table 1

The Number of Data Points in the Treatment Phase and the Corresponding Number of Data Points That Must Be Above Both Criterion Lines to Conclude That There is a Reliable Treatment Effect Using the DC or CDC Method

Treatment phase	Needed above both criterion lines
5	5
6	6
7	6
8	7
9	8
10	8
11	9
12	9
13	10
14	11
15	12
16	12
17	12
18	13
19	13
20	14
21	14
22	15
23	15

treatment. Thus, participants were presented with an example of a correct “yes” response and a correct “no” response. The experimenter modeled and verbally explained how to apply the rules to each example graph. Training lasted from 10 to 15 min.

Treatment. Procedures in treatment were identical to those used in baseline. However, in treatment, participants had been exposed to training and also judged graphs that included the DC criterion lines.

RESULTS AND DISCUSSION

Results for all 5 participants are shown in Figure 3. Baseline and treatment data were similar for all participants. In baseline, the percentage of correct interpretations averaged 51.7%, 47.5%, 57.0%, 65.0%, and 55.8%, for Participants 1, 2, 3, 4, and 5, respectively. In treatment, the percentage of correct interpretations averaged 95.0%, 88.3%, 93.8%, 93.0%, and 97.5%, for Par-

ticipants 1, 2, 3, 4, and 5, respectively. The results show that individuals can be rapidly trained to interpret A-B single-case designs.

We reviewed the errors made by the participants and also asked them what difficulties they experienced when they implemented the DC method. Almost all the errors were ones in which one or more of the data points overlapped with one of the criterion lines such that it was difficult to determine whether the center point of the data point fell above or below the criterion line. Verbal reports from the participants also indicated that this was a limitation of the DC method. One way to mitigate the impact of this limitation would be to instruct visual inspectors not to count data points as meeting the criterion whenever there is uncertainty as to whether the point is above the criterion line (i.e., to err on the side of caution). This would make human implementation of the DC method slightly more conservative than the computer implementation conducted in Study 1. An alternative way to mitigate the effects of this limitation of the DC method would be to train the visual inspectors to examine the actual number in the worksheet corresponding to the data point in question and to compare it with the number of the corresponding point for the relevant criterion line.

We also examined participant errors to determine whether the participants tended to make proportionally more Type I or more Type II errors during baseline, and whether implementation of the DC method altered the relative percentages of Type I and Type II errors. Across participants, the mean percentages of Type I errors ($M = 23.7\%$; range, 11.6% to 30%) and Type II errors ($M = 20.2\%$; range, 16% to 36.7%) were similar in baseline. During treatment, Type I errors were almost eliminated ($M = 0.25\%$; only 1 participant made one Type I error out of 400 interpretations) and Type II errors were markedly reduced ($M = 6.0\%$; range, 2.5% to 11.7%). Thus, the DC method not only

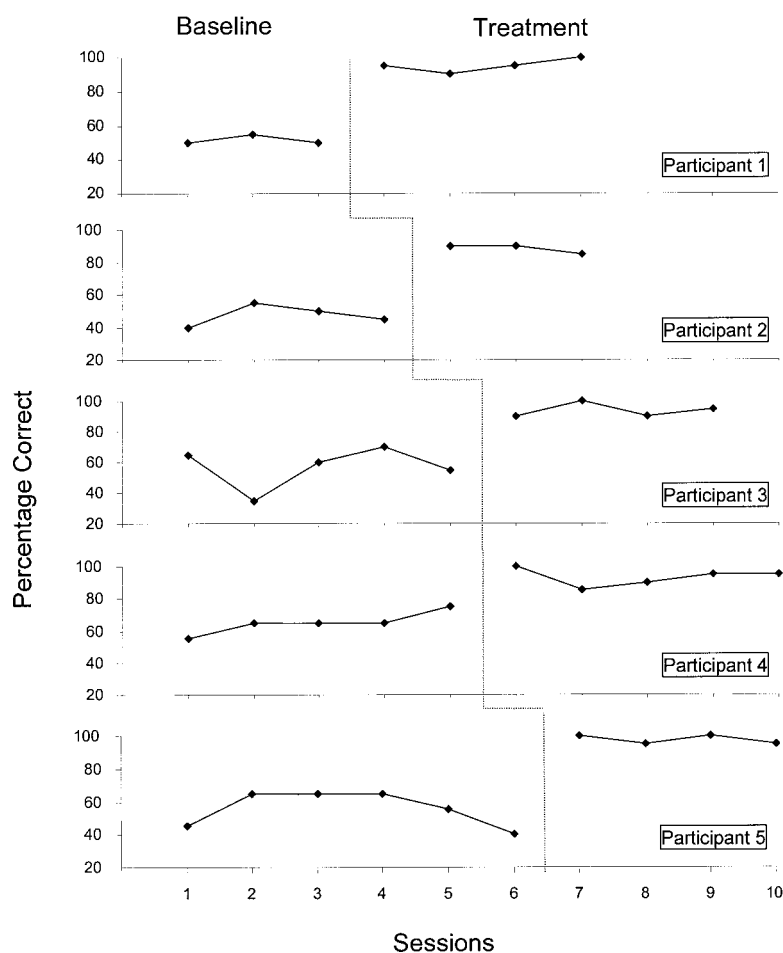


Figure 3. The percentages of correct interpretations during baseline and treatment for the 5 participants in Study 2.

reduced the total number of visual inspection errors but it also shifted the inspectors' biases toward more conservative types of errors. Whereas Type I errors accounted for 54% of the errors made across participants during baseline, only 1 of 25 errors (4%) made in treatment was a Type I error.

STUDY 3: GROUP TRAINING OF VISUAL INSPECTORS IN THE DC METHOD

The purpose of Study 3 was to determine whether the training methods used in Study 2 could be incorporated into a PowerPoint™

presentation and used to rapidly train large groups of visual inspectors to interpret A-B graphs using the DC method. As mentioned above, we would have trained the participants in the CDC method had the results of Study 1 been available at the outset of Study 3. Nevertheless, the target responses required of the visual inspectors are identical for the DC and CDC methods; thus, the results of Study 3 should readily generalize to the CDC method.

METHOD

Materials

Three 20-graph packets were constructed using methods identical to those described

in Study 2 except that graphs with obvious treatment effects were not eliminated from the packets. However, the graphs were not printed. Rather, the graphs were in a PowerPoint[™] presentation format and were projected onto a screen at the front of the room using a Toshiba LCD projector (Model TLP 260). The participants recorded their interpretations of the graphs on answer sheets that were similar to the ones used in Study 2 except that each answer sheet had either a large A or a large B at the top of the answer sheet. The order of A and B answer sheets was randomized so that participants would be randomly assigned to either Group A or Group B when the answer sheets were distributed.

Participants and Setting

Participants were 87 adults attending a workshop on behavior analysis at an annual meeting of a state chapter of the Association for Behavior Analysis. Participants were seated at long tables in about 10 rows, which could accommodate about 15 individuals per row. Participants who received an answer sheet with a large A at the top were assigned to Group A; those who received an answer sheet with a large B at the top were assigned to Group B.

Response Measurement and Reliability

Participants recorded their interpretations on a preprinted answer sheet by circling the word "yes" if they judged that there was a reliable treatment effect and "no" if they judged otherwise. An experimenter scored the accuracy of responses, and a second experimenter independently scored 33% of the answer sheets to ascertain the reliability of the scoring method, which was 100%.

Procedure

The visual inspection assessment and training was a small component of the workshop. This component started around 10:00

a.m., just prior to the first scheduled break. Prior to the start of the first session, the experimenter explained how to mark answers on the data sheet and told the participants to mark "yes" whenever they judged a treatment effect to be reliable, even if it was a small but reliable effect. The participants were told that being able to detect small but reliable treatment effects is sometimes important, and they were given the example that small decreases in mean blood pressure can produce clinically important health benefits. The participants had unlimited time to judge each graph, but sessions generally lasted about 5 to 15 min.

Initial baseline session: Both groups. The experimenter projected the first set of 20 graphs on the screen, one at a time, and asked the participants in both groups to interpret them. The baseline graphs were similar in form to the one depicted in the top panel of Figure 1 (except that they were projected onto a screen rather than printed on paper). When the participants finished this initial baseline session, the participants in Group B were dismissed for their break.

Training: Group A. While Group B was on break, Group A received training in the DC method. The training was identical to the procedures used in Study 2 except that the example graphs were projected onto the screen at the front of the room rather than printed on paper. Written instructions were provided on the answer sheet. Training lasted approximately 10 to 15 min.

Treatment Session 1: Group A. Immediately following training, participants in Group A completed a treatment session. Procedures in treatment were identical to those used in baseline. However, in treatment, participants had been exposed to training and the graphs projected on the screen included the DC criterion lines. After Group A completed this treatment session, they were dismissed for their break.

Baseline Session 2: Group B. While Group A was on break, participants in Group B re-

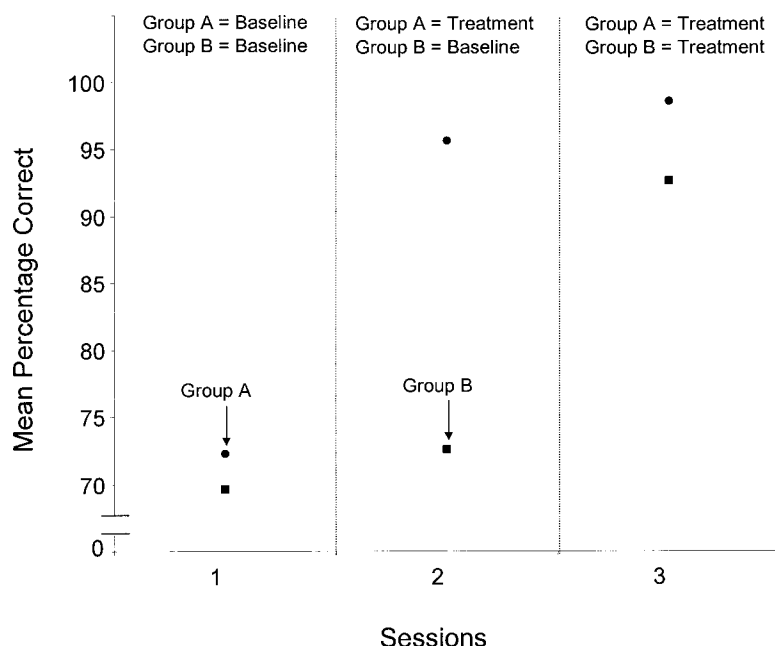


Figure 4. The percentages of correct interpretations for Groups A and B in baseline and treatment for the 87 participants in Study 3.

turned and completed a second baseline session. The second baseline session was identical to the first (except that Group A was not present).

Training: Group B. Immediately following this second baseline, Group B received training in the DC method. The training was identical to the procedures used to train Group A. Training lasted approximately 10 to 15 min.

Final treatment session: Both groups. Immediately following training for Group B, Group A returned from their break and participants in both groups completed the final treatment session. Procedures in this treatment session were identical to those described above for Treatment Session 1 (except that both groups were present).

RESULTS AND DISCUSSION

Results of Study 3 are shown in Figure 4. During the initial baseline, both groups produced correct interpretations ($M_s = 72\%$ and 70% for Groups A and B, respectively) at levels somewhat above chance (i.e., 50%), but

the performance of the two groups during baseline was not significantly different, $F(1, 85) = 1.13$, $p = .29$. During the second session (after Group A was trained in the DC method and Group B was not), the percentage of correct interpretations increased markedly for Group A ($M = 96\%$), whereas the increase for Group B ($M = 73\%$) was marginal. In the final treatment session, after both groups had been trained in the DC method, the level of correct interpretations was maintained at high levels for Group A ($M = 98\%$) and increased markedly for Group B ($M = 93\%$). Across both groups, the percentage of correct interpretations increased from an initial baseline mean of 71% to a final treatment mean of 95% . A repeated measures analysis of variance was conducted with one between-groups factor (Groups A and B) and one within-subject factor for time of the session (Sessions 1, 2, and 3). As expected, there was a significant effect for group membership, $F(1, 260) = 39.78$, $p < .0001$, a significant effect for time of session, $F(2, 260) = 158.98$, $p < .0001$,

and a significant interaction between group membership and time of session, $F(2, 260) = 31.45$, $p < .0001$. In addition, whereas only 1 individual interpreted the initial baseline graphs at mastery level (90% or more correct), 73 (84%) of the participants attained mastery level during the final treatment session. Thus, with a training time of about 15 min, a large group of participants was trained in the DC method, marked improvements occurred in the mean level of correct interpretations, and most of the participants attained mastery performance. Results of Study 3 indicate that these training methods could easily be incorporated into a staff orientation or training program to rapidly train new staff to interpret A-B designs.

GENERAL DISCUSSION

Study 1 answered basic questions about the accuracy of several methods of interpreting A-B designs. Study 2 showed how the information from Study 1 could be used to increase the accuracy of human visual inspectors and to bias the few errors they made following training toward greater conservatism (i.e., decreasing the ratio of Type I to Type II errors). Finally, Study 3 showed how the training procedures in Study 2 could be incorporated into a format that would facilitate rapid training of large groups of individuals to interpret single-case designs. This training format could easily be incorporated into an ongoing staff orientation or training program.

The results of Study 1 add to the literature on visual and statistical interpretation of single-case designs in several ways. First, they showed that two relatively simple refinements of the SM method of visual inspection, called the DC and CDC methods, could perform as well or better than more complex statistical analyses like the GLM and ITSE. In fact, the CDC was the only method, visual or statistical, that consistently guarded against Type I errors, and it did so

while producing better power than either the GLM or ITSE. The reasons for this finding are not entirely clear. We speculate that the GLM and ITSE methods may have reduced power (relative to the CDC method) because they each require multiple predictor terms in the full model, which reduces the number of degrees of freedom in the denominator of the F test.

A number of proponents of statistical analysis as an alternative or adjunct to visual inspection have suggested that when a single-case series appears to be serially dependent, it may be advisable to use a statistical method that includes one or more autoregressive parameters (i.e., that it may be important to model the level of serial dependence in the data series statistically; Crosbie, 1995; Gorman & Allison, 1996; Gottman, 1981; Robey, Schultz, Crawford, & Sinner, 1999). Our results suggest that there may be simpler ways to protect against the biasing effects of serial dependence, at least in the range of autocorrelation tested in the current investigation (which exceeded the highest level of autocorrelation [0.47] reported in the Huitema, 1985, survey). That is, we obtained reasonable protection against Type I errors resulting from first-order autocorrelation (while maintaining reasonable levels of power) simply by raising the heights of the two DC criterion lines to create the CDC method.

Results from Study 2 suggest that the rates of Type I errors for visual inspection (16% to 84%) reported by Matyas and Greenwood (1990) appear to have been inflated. Matyas and Greenwood reported that visual inspectors made Type I errors at higher rates than Type II errors. In the introduction of the current article, we questioned whether at least some of the high Type I error rates reported in the Matyas and Greenwood investigation could have been due to sampling error (i.e., that only one graph was generated from each set of autoregressive parameters, and in some cases, the one graph may not have been repre-

sentative of the autoregressive model from which it was drawn). Results from Study 2 are consistent with this hypothesis in that, during baseline, visual inspectors made Type I errors ($M = 23\%$) at considerably lower levels than those reported by Matyas and Greenwood and at about the same rate as they made Type II errors ($M = 20\%$). More important, once the visual inspectors were trained in the DC method, Type I errors were almost eliminated and Type II errors were also markedly reduced.

It should be noted that no graphs were included in Study 2 in which the correct implementation of the DC method resulted in an error, because the main purpose of Study 2 was to evaluate the training methods we employed. That is, precise implementation of the DC and CDC methods by visual inspectors over a large number of data sets would be expected to result in errors close to those reported in Study 1.

We have presented the SM, DC, and CDC methods as visual inspection procedures, but one might reasonably question whether they represent hybrids of visual inspection and statistical methods, because the process begins with visual inspection but then applies a numerical criterion (SM method) or two criteria (DC and CDC methods) originally based on the binomial test. However, we developed the DC method through visual inspection and interpretation of the pattern of errors made by the SM method. Moreover, we did not use the binomial test in a manner consistent with its statistical properties or assumptions. We simply used the numbers from the binomial test as a starting point and then empirically tested the performance of the DC method using Monte Carlo simulation, which led to the creation of the CDC method. Certainly, we applied statistical methods as a part of the Monte Carlo simulation to test and refine the methods. However, the inclusion of the values from the binomial test was not critical to the develop-

ment or refinement of these interpretive methods. We could have alternatively generated the original numerical criteria shown in Table 1 by having an expert panel of visual inspectors sort the graphs into categories (e.g., treatment effect present, no treatment effect) and then setting the criteria so that the DC method closely approximated the decisions of the panel, as was done in the Hagogian *et al.* (1997) investigation.

The results of Study 1 should be interpreted relative to a number of limitations. First, the CDC method does not take into account the magnitude of a treatment effect. For example, implementation of extinction might result in a clinically significant reduction in problem behavior, but only after an initial increase in the response (i.e., an extinction burst). An experienced visual inspector would probably detect this treatment effect much sooner than the CDC criteria. Second, we did not test Type I and Type II error rates for autoregressive models with negative autocorrelation or with positive autocorrelation values above 0.5. We did not include negative autocorrelation values because negative autocorrelation tends to decrease Type I error rates. We did not include higher autocorrelation values partly because observed autocorrelation values generally do not exceed 0.5 in single-case data series (Huitema, 1985), but mostly because our autoregressive model would not produce observed first-order autocorrelation coefficients above 0.5 (e.g., an input value of $a = 0.6$ would produce an observed autocorrelation value similar to an input value of $a = 0.4$). We chose this autoregressive model because it produced an observed autocorrelation that was reasonably close to the input value we entered and because it allowed us to vary effect size in a linear fashion. Future research should be directed toward comparing visual and statistical interpretive methods using a broader range of autocorrelation values, including negative values.

Another limitation of the series of Studies 1 through 3 is that the CDC method was shown to produce the best protection against Type I errors in combination with reasonable power to detect real treatment effects in Study 1, but the participants in Studies 2 and 3 were trained in the DC method. This seems to be a minor limitation, at most, because the target behaviors required of the visual inspectors are identical for the DC and CDC methods, unless the visual inspectors generate the criterion lines by hand, which was not the case in either Study 1 or Study 2.

Another inconsistency between Study 1 and Studies 2 and 3 is the fact that a computer implemented the SM, DC, and CDC methods for the Monte Carlo simulation in Study 1, whereas human visual inspectors implemented the DC method in Studies 2 and 3. We used a computer to simulate the visual inspection process involved in these three interpretive methods for practical reasons (i.e., because 600,000 simulated data sets were interpreted in Study 1). Nevertheless, this raises the question as to why not just have a computer implement the CDC method instead of a human visual inspector?

We would caution against a purely mechanical implementation of the CDC method by a computer or by a human visual inspector. We developed the CDC method as a tool to rapidly jump-start the training of new staff in visual inspection procedures, and the results of Study 3 demonstrate its potential for this purpose. But we do not believe that mastery-level implementation of the CDC method should be an endpoint for training in visual inspection methods; rather, it has the potential to be a useful first step. In addition, we attempted to make interpretation as easy as possible by providing graphs with the criterion lines drawn on them. The process of interpreting the graphs would obviously be more difficult if the new staff had to generate the criterion lines themselves. Therefore, we developed an Excel[™] spreadsheet that does almost all of the

work for the user. The user enters the raw data and the spreadsheet creates the graph with the criterion lines and calculates (a) the number of treatment data points, (b) the number of treatment data points that fall above both criterion lines (or below, for behavior reduction), and (c) the number of data points that need to be above both criterion lines in order to conclude that there is a reliable treatment effect. This worksheet (called Visual Inspect AB.xlt) along with a set of instructions can be downloaded free of charge at www.marcus.org/fisher.

The CDC method does not assist the visual inspector in identifying a number of important behavioral phenomena (e.g., extinction bursts, behavioral contrast), nor does it aid in determining whether a treatment effect that is judged to be reliable actually represents a socially meaningful change in the behavior of interest. Such judgments require that behavior analysts understand the context in which the clinical or scientific question is being posed. It is possible that the CDC method will detect a small but reliable treatment effect that is not at all clinically significant. Behavior analysts not only need to remain in close contact with their data but also with their consumers in order to make reasonable judgments about the clinical relevance of treatment effects.

REFERENCES

- Allison, D. B., & Gorman, B. S. (1993). POWCOR: A power analysis and sample size program for testing differences between dependent and independent correlations. *Educational & Psychological Measurement, 53*, 133–137.
- Baer, D. M. (1977). "Perhaps it would be better not to know everything." *Journal of Applied Behavior Analysis, 10*, 167–172.
- Bailey, D. B. (1984). Effects of lines of progress and semilogarithmic charts on ratings of charted data. *Journal of Applied Behavior Analysis, 17*, 359–365.
- Boykin, R. A., & Nelson, R. O. (1981). The effects of instructions and calculation procedures on observers' accuracy, agreement, and calculation cor-

- rectness. *Journal of Applied Behavior Analysis*, 14, 479–489.
- Center, B. A., Skiba, R. J., & Casey, A. (1985–1986). A methodology for the quantitative synthesis of intra-subject design research. *Journal of Special Education*, 19, 387–400.
- Crosbie, J. (1987). The inability of the binomial test to control Type I error with single-subject data. *Behavioral Assessment*, 9, 141–150.
- Crosbie, J. (1995). Interrupted time-series analysis with short series: Why it is problematic; how it can be improved. In J. M. Gottman & G. Sackett (Eds.), *The analysis of change* (pp. 361–395). Mahwah, NJ: Erlbaum.
- DeProspero, A., & Cohen, S. (1979). Inconsistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis*, 12, 573–579.
- Ferron, J. M., & Ware, W. B. (1995). Analyzing single-case data: The power of randomization tests. *Journal of Experimental Education*, 63, 167–178.
- Franklin, R. D., Gorman, B. S., Beasley, T. M., & Allison, D. B. (1996). Graphical display and visual analysis. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single case research* (pp. 119–158). Mahwah, NJ: Erlbaum.
- Gorman, B. S., & Allison, D. B. (1996). Statistical alternatives. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single case research* (pp. 159–214). Mahwah, NJ: Erlbaum.
- Gottman, J. M. (1981). *Time-series analysis: A comprehensive introduction for social scientists*. Cambridge: Cambridge University Press.
- Hagopian, L. P., Fisher, W. W., Thompson, R. H., Owen-DeSchryver, J., Iwata, B. A., & Wacker, D. P. (1997). Toward the development of structured criteria for interpretation of functional analysis data. *Journal of Applied Behavior Analysis*, 30, 313–326.
- Harbst, K. B., Ottenbacher, K. J., & Harris, S. R. (1991). Interrater reliability of therapists' judgments of graphed data. *Physical Therapy*, 72, 107–115.
- Huitema, B. E. (1985). Autocorrelation in applied behavior analysis: A myth. *Behavioral Assessment*, 7, 107–118.
- 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.
- Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings*. New York: Oxford University Press.
- 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.
- Matyas, T. A., & Greenwood, K. M. (1996). Serial dependency in single-case time series. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single case research* (pp. 215–243). Mahwah, NJ: Erlbaum.
- Michael, J. (1974). Statistical inference for individual organism research: Mixed blessing or curse? *Journal of Applied Behavior Analysis*, 7, 647–653.
- Ottenbacher, K. J. (1990). Visual inspection of single-subject data: An empirical analysis. *Mental Retardation*, 28, 283–290.
- Page, T. J., Iwata, B. A., & Reid, D. H. (1982). Pyramidal training: A large-scale application with institutional staff. *Journal of Applied Behavior Analysis*, 15, 335–351.
- Park, H., Marascuilo, L., & Gaylord-Ross, R. (1990). Visual inspection and statistical analysis in single-case designs. *Journal of Experimental Education*, 58, 311–320.
- Parsonson, B. S., & Baer, D. M. (1986). The graphic analysis of data. In A. Poling & R. W. Fuqua (Eds.), *Research methods in applied behavior analysis: Issues and advances* (pp. 157–186). New York: Plenum.
- Reid, D. H., & Parsons, M. B. (1995). Comparing choice and questionnaire measures of the acceptability of a staff training procedure. *Journal of Applied Behavior Analysis*, 28, 95–96.
- Robey, R. R., Schultz, M. C., Crawford, A. B., & Sinner, C. A. (1999). Single subject clinical outcome research: Design, data, effect sizes, and analyses. *Aphasiology*, 13, 445–473.
- Rojahn, J., & Schulze, H. H. (1985). The linear regression line as a judgmental aid in visual analysis of serially dependent A-B time-series data. *Journal of Psychopathology & Behavioral Assessment*, 7, 191–206.
- Watson, T. S., & Kramer, J. J. (1995). Teaching problem solving skills to teachers-in-training: An analogue experimental analysis of three methods. *Journal of Behavioral Education*, 5, 281–293.
- White, O. R. (1974). *The "split middle"—a "quickie" method of trend estimation*. Seattle: Experimental Education Unit, Child Development and Mental Retardation Center, University of Washington.

Received February 13, 2003

Final acceptance May 23, 2003

Action Editor, David P. Wacker