On Methods

Visual Inspection of Data Revisited: Do the Eyes Still Have It?

Gene S. Fisch
Yale University

In behavior analysis, visual inspection of graphic information is the standard by which data are evaluated. Efforts to supplement visual inspection using inferential statistical procedures to assess intervention effects (e.g., analysis of variance or time-series analysis) have met with opposition. However, when serial dependence is present in the data, the use of visual inspection by itself may prove to be problematic. Previously published reports demonstrate that autocorrelated data influence trained observers' ability to identify level treatment effects and trends that occur in the intervention phase of experiments. In this report, four recent studies are presented in which autoregressive equations were used to produce point-to-point functions to simulate experimental data. In each study, various parameters were manipulated to assess trained observers' responses to changes in point-to-point functions from the baseline condition to intervention. Level shifts over baseline behavior (treatment effect), as well as no change from baseline (no treatment effect or trend), were most readily identified by observers, but trends were rarely recognized. Furthermore, other factors previously thought to augment and improve observers' responses had no impact. Results are discussed in terms of the use of visual inspection and the training of behavior analysts.

Key words: visual inspection, data analysis, statistical analysis

Although classical graphic methods were developed more than a century earlier (cf. Spence & Lewandowsky, 1990), visual inspection of data as a technique applied to behavior-analytic phenomena has its roots in the concluding chapter of *The Behavior of Organisms* (Skinner, 1938). Skinner states that, "Where a reasonable degree of smoothness and reproducibility can be obtained with a few cases . . . there is little reason . . . to consider large numbers" (p. 442), and "In the simple sense of involving large numbers of measurements, very little of the preceding work is statistical" (p. 442). He states further that

> The statistical approach is characterized by relatively unrefined methods of measurement and a general neglect of the problem of direct description; [whereas] the non-statistical approach confines itself to specific instances of behavior and to the development of methods of direct measurement and analysis. (p. 443)

To emphasize his belief that a science of behavior is about single organisms in which statistical analysis has no place, Skinner cites as an example the case of a physician who is"trying to determine whether his patient will die before morning. . . . [He] make[s] little use of actuarial tables. Nor can the student of behavior predict what a single organism can do if his laws only apply to groups" (1938, p. 443). However, Skinner goes to say that,

> It may be that the differences between the two approaches [statistical and nonstatistical] are transitory and that eventually a combination of the two will give us our best methods, but at the present time they are . . . different and almost incompatible conceptions of a science of behavior. (p. 443)

That difference, calculating averages of noisy group data to represent individuals as opposed to examining the behavior of

I thank three anonymous reviewers for their many comments and suggestions. Portions of this paper were previously presented at the annual meetings of the Association for Behavior Analysis, Chicago, 1993, the joint meetings of the American Statistical Association, Boston, 1992, and San Francisco, 1993, and the International Society for Optical Engineering, San Jose, 1994.

Correspondence can be addressed to Gene S. Fisch, Yale University General Clinical Research Center, 530 Hunter Building, 333 Cedar St., New Haven, Connecticut 06520.

individual organisms themselves, was underscored by Sidman (1952). He identified serious methodological weaknesses concerning the validity of inferences derived from averaged data to describe a functional relationship about individuals, from which Bakan (1954) later provided a more generalized argument. Despite Estes' (1956) contention that a "valid interpretation of group curves depends on principles in common in all problems of statistical inference" (p. 139), the rift between statistical and nonstatistical approaches to a science of behavior widened.

The contrast between the two approaches was emphasized by Sidman (1960):

> A statistical judgment of significance or non-significance may itself be the product of chance. . . . [and] what is meant by "chance"? To some . . . chance is simply a name for the combined effect of uncontrolled variables. If such variables are, in fact, controllable, then chance is simply an excuse for sloppy experimentation, and no further comment is required. (p. 45)

Inferential statistics were no longer considered a research tool for psychologists but a marker for scientific sloth.

As long as visual inspection of graphed data was confined to the analysis of behavior involved primarily with nonhuman subjects under tight experimental control, there was little argument from the increasing ranks of experimental psychologists engaged in such research. However, as the principles of behavioral analysis spread to areas of investigation beyond the laboratory and involved human subjects whose histories were largely unknown, were studied under less-than-ideal experimental circumstances, and who presented a broad array of problematic behavioral repertoires, complications developed.

One problem arose from the use of human observers to record data that had previously been obtained from electromechanical devices. As measuring instruments, observers differ from one another in many ways. Among them are history, training, experience, and vigilance. Hence, interobserver or interrater reliability became an important issue in applied behavior analysis. In its 10th anniversary volume, the *Journal of Applied Behavior Analysis* (*JABA*) included a report by Kelly (1977) that reviewed 8 years of experimental data from observations of human behavior published by the journal. Although 94% of reports contained information on interrater reliability, the vast majority of them attained less than 90% agreement between observers. Also in that issue, several authors illustrated how percentage agreement was inadequate as a measure of agreement. Yeltman, Wildman, and Erickson (1977) showed that percentage agreement was affected by behavior frequency and by whether nonoccurrences of behavior were included in the measure. Kratochwill and Wetzel (1977) noted that percentage agreement may be insensitive to certain response definitions. Kazdin (1977) argued further that, among other factors, observer drift and complexity of coding could affect reliability and recommended measures to improve interobserver agreement. Hartmann (1977) suggested other ways to compute interobserver reliability. Kratochwill and Wetzel (1977) and Yeltman et al. (1977) provided additional statistics for computing interrater agreement that would correct for chance agreement. Suggestions regarding the use of such statistical procedures to compute interobserver reliability were also met with some skepticism (Baer, 1977).

This was not the first attempt to introduce inferential statistics to an applied behavioral setting. Previously, Gentile, Roden, and Klein (1972) proposed the use of analysis of variance (ANOVA) to evaluate data collected from single-subject studies. One assumption made by Gentile et al., that successive observations within treatment conditions be considered as independent events, is essential in using ANOVA procedures. Subsequent rebuttals of Gentile et al. were broadly critical of the approach.

Some were troubled by the particular ANOVA model (Kratochwill et al., 1974), whereas others were more concerned about using ANOVA on the serially dependent data that commonly occur in behavioral analysis (Keselman & Leventhal, 1974; Thoresen & Elashoff, 1974; Toothaker, Banz, Noble, Camp, & Davis, 1983). Serially dependent data, which are also referred to as serially correlated or autocorrelated data, denote a temporally ordered series of events in which a measurement ascertained in one time period is related to (i.e., depends on) a value or values obtained earlier. Keselman and Leventhal (1974), Thoresen and Elashoff (1974), and Toothaker et al. (1983) suggested using time-series analyses (TSA) that routinely include the assumption of serial dependence in their models.[1] As with interrater agreement, there was some disquiet about any use of inferential statistics (Michael, 1974).

In statistical analysis, serial dependence is problematic, because its occurrence violates several important assumptions. Ordinarily, events are assumed to be independent from another so that differences from a sample mean occur in a normally distributed fashion. When two sample means are compared, serial correlation will inflate the probability of falsely rejecting the null hypothesis, as Crosbie (1987) and Toothaker et al. (1983) have shown. For visual inspection, moderate positive autocorrelation, as opposed to a level shift in responding during the intervention phase, may produce a modest increasing trend. Given session-to-session response variability, the trend could be missed and the modest effect obscured. Consequently, a variable would likely be dropped prematurely from the study (DeProspero & Cohen, 1979).

It should come as no great surprise that behavioral data from single-subject design studies exhibit serial dependence and that responses from one session are related to those from the next. In this science of behavior, there would otherwise be no prospect of maintaining previously trained responses, demonstrating their extinction, or illustrating the acquisition of novel behaviors. Jones, Vaught, and Weinrott (1977) highlighted the point as they incorporated TSA into operant research. They examined both level shifts in responding and changes in trends from baseline to intervention as manifested in single-subject session-to-session data (point-to-point functions) presented in previously published studies, thereby investigating behavioral data along a dimension different from either visual inspection or frequentist statistics. By using TSA, they hoped to supplement rather than replace visual inspection as a tool for evaluating data. Hartmann et al. (1980) suggested that ITSA be used in studies in which the experimental effect may be small and serial correlation is present in the data.

Because serial dependence posed a serious problem for statistical analysis, Jones, Weinrott, and Vaught (1978) designed another study to determine whether serial correlation was also an obstacle to the visual inspection of data. Published experiments from *JABA* were selected in which significant within-phase autocorrelations were found. Graphs of the data were shown to 11 judges who were then asked whether a "meaningful" (i.e., reliable) change had been demonstrated from one phase to the next in the graphs shown. The extent to which ob-

[1] Time-series analyses are statistical procedures adapted from physics by which autocorrelated components of a function can be identified and systematically transformed so that data can be subsequently evaluated using general linear methods (e.g., regression analysis). In general, TSA requires many values to analyze data from baseline and intervention phases, although interrupted TSA (ITSA) procedures attempt to use fewer points. More detailed discussions of TSA, ITSA, and alternatives to ITSA can be found in Glass, Willson, and Gottman (1975), Velicer and Harrop (1983), Harrop and Velicer (1985), Greenwood and Matyas (1990), Matyas and Greenwood (1991), and Crosbie (1993).

servers agreed that a significant treatment effect was present was inversely related to the degree of autocorrelation in the point-to-point functions presented. Moreover, except for cases in which both statistical significance and serial dependence were low, average agreement among judges was modest, ranging from .48 to .60. In a similar study, DeProspero and Cohen (1979) constructed 36 ABAB reversal design graphs using 10 points per phase and asked raters to indicate the degree to which experimental control had been demonstrated. Mean correlation among observers was .61, comparable to the level of agreement obtained by Jones et al. (1978). Later, Wampold and Furlong (1981) and Furlong and Wampold (1982) found that not only do observers focus primarily on the size of the treatment effect, but they are unable to differentiate the treatment effect clearly and consistently from the trend. Thus, serial dependence also posed a serious problem for visual inspection.

The twin issues of autocorrelation and the application of TSA to temporally based single-subject data were examined in detail by Huitema (1986). Using an elegant example, Huitema argued that the conceptualization of autocorrelation with raw data, as opposed to the residuals (i.e., the difference between observed values and their estimates obtained from regression analysis) described by Jones et al. (1977), was logically flawed. Moreover, calculation of autocorrelation in residuals must be executed on separate design phases. Otherwise, autocorrelation would be detected where none in fact existed. Huitema also presented results he had evaluated previously (Huitema, 1985) in which the average autocorrelation coefficent from 441 data sets was nearly zero, and concluded that serial dependence in applied behavior analysis was a myth.

Huitema's (1986) conclusions regarding the existence (or lack thereof) of autocorrelation in applied behavioral studies were taken to task by several investigators. Busk and Marascuilo (1988) reevaluated the studies examined by Huitema (1985) and, after correcting for sample-size differences, found that 40% of baseline phases and 59% of intervention phases displayed autocorrelation coefficients greater than .25. These authors calculated that autocorrelation coefficients of that magnitude would inflate Type I errors by more than 100%. Sharpley and Alavosius (1988) noted further that examining residuals or raw data was irrelevant regarding the effect of autocorrelation on statistical inference. (For a more complete discussion of autocorrelation calculated from raw data compared to residuals, see Busk & Marascuilo, 1988; Huitema, 1986; Matyas & Greenwood, 1991.) Matyas and Greenwood demonstrated that Lag 1 autocorrelations exist in applied behavior analysis studies but statistical procedures used to assess them would be problematic when the number of points per phase was small.

These findings indicate that serial dependence in applied behavior analysis is not of minor significance. More important, when autocorrelation occurs, can it be identified with, and possibly managed by, statistical techniques? Or, perhaps more meaningfully to traditional behavior analysts, when it occurs, how does it affect the visual inspection of graphic data?

THE USE OF LAG 1 AUTOCORRELATION EQUATIONS TO DETERMINE THE EFFECT OF SERIAL CORRELATION ON VISUAL INSPECTION

Matyas and Greenwood (1990) generated graphs constructed from Lag 1 autocorrelated equations to examine trained observers' abilities to detect level treatment effects or trends manifested in point-to-point functions. They used the following equation:

$$Y_t = aY_{t-1} + b + d + e, \qquad (1)$$

where Y_t is the ordinate at session t, Y_{t-1} is the ordinate at session $t - 1$, a is the autoregressive component used to introduce a trend, b is the ordinate at $t = 0$, d is the size of the treatment effect, and e introduces variability into the function using randomly selected values from a normal distribution.[2] Matyas and Greenwood (1990) asked subjects to identify the response that best described the function in the intervention condition only: If they observed a level change only from the baseline value in the intervention phase (i.e., a level treatment effect); if they observed a change in trend only from the baseline to intervention phase (i.e., trend alone); if they observed a level treatment effect and a change in trend; if they observed neither a level treatment effect nor trend; or if they observed some other systematic change from baseline to treatment. In studies that examine the effects of different types of intervention, change in trend has been defined generally as a change in the angle or pitch in slope (usually positive) from a baseline with zero slope (Hartmann et al., 1980; Ottenbacher, 1986; Wampold & Furlong, 1981). Although a trend implies positive autocorrelation, the reverse may not be true (Crosbie, 1989). However, positive autocorrelation with moderate variability will create a point-to-point function that appears as a trend or change in trend (Matyas & Greenwood, 1990).

Matyas and Greenwood (1990) were interested in evaluating the impact of autocorrelation on Type I and Type II errors. In statistics, a Type I error occurs when the null hypothesis is rejected but is true. A Type II error occurs when the alternate hypothesis is rejected but is true. More commonly in psychology and in signal-detection theory in particular, Type I errors refer to false alarms, and Type II errors refer to misses.

To assess the influence of autocorrelation on errors, Matyas and Greenwood (1990) manipulated several parameters; specifically, the value of the autoregression component ($a = 0.0$, 0.3, 0.6), the magnitude of the intervention effect ($d = 0$, 5, 10), and the amount of variability ($s = 1$, 3, 5). The value of e was calculated by selecting numbers from a table of random digits, the relative frequencies of which follow a normal distribution with mean and standard deviation (0, 1), then multiplied by a constant (s) to magnify variability. Not surprisingly, results indicated that treatment effects were easier to detect when variability was low and effect magnitude was large. On the other hand, when serial correlation was present, subjects often identified the trend produced by the autoregressive component as if it were a treatment effect.

Their results, along with those obtained by Cleveland and McGill (1984, 1985, 1987), Lewandowsky and Spence (1989), Spence (1990), and Spence and Lewandowsky (1990), have important implications for the visual inspection of graphed data. Consequently, Fisch and his colleagues attempted a systematic replication of the results obtained by Matyas and Greenwood (1990). Contingent upon replicating their results, Fisch sought to explore additional parameters that might affect visual inspection.

Procedurally, the studies by Fisch grew out of the methodology employed by Matyas and Greenwood (1990). Subjects were presented with a set of graphs constructed from a first order autoregression equation, as in Equation

[2] Lag 1 autocorrelated functions can be partitioned into three basic types, according to the value of the autoregressive component a. When $a = 0$, there is no autocorrelation and Y_t is a function of its initial value, b, and the variability produced by e. When e is small, the function appears much like a horizontal line. When $a < 0$, negative values of a produce alternating negative and positive values for Y_{t-1}. If $a = -1$, the autoregressive function appears as a sawtooth pattern. When $a > 0$, positive values produce a possibly increasing trend in the autocorrelated function, depending on the magnitiude of a and the initial value, b. A more complete description of autocorrelation in behavioral research can be found in Huitema (1986).

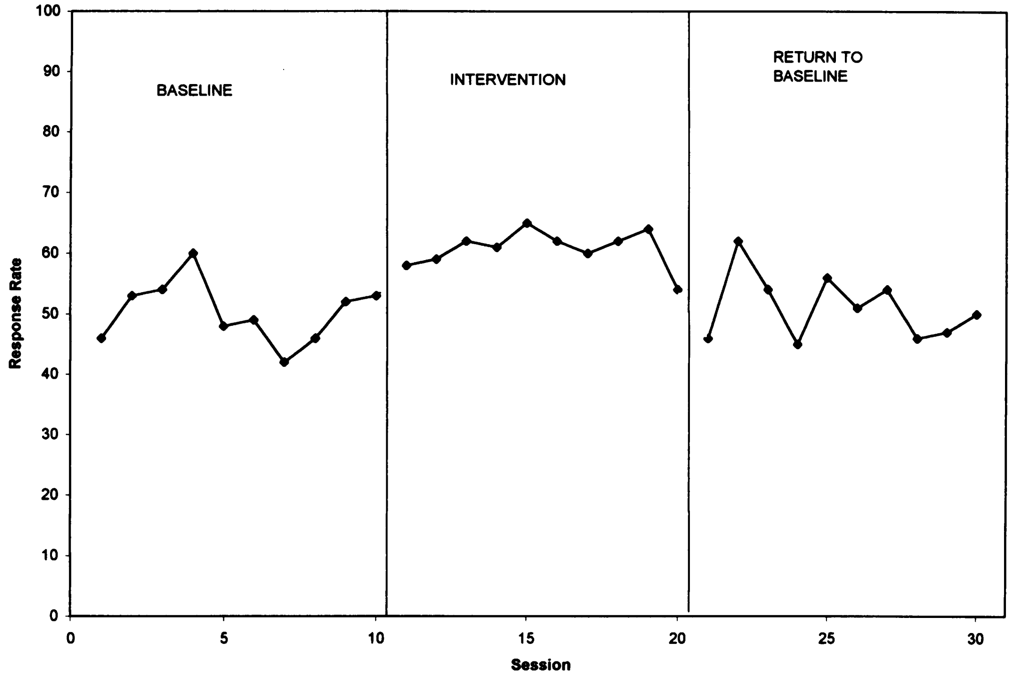

Figure 1. Sample graph of point-to-point functions generated by an autoregressive equation.

1. Each graph contained at least a baseline and an intervention phase. Subjects were given response sheets containing the letters A, B, C, D, and E for each graph and were asked to examine the figures one at a time. They were told to circle one of the letters: A if they saw a level shift from the baseline to intervention phase, B if they observed a change in trend from baseline to intervention, C if they noted both a level shift and a trend, D if they saw no change from baseline to intervention, and E if there was some other systematic change. No time constraints were imposed, but subjects were asked to work as quickly as possible. An example of the type of graphs is shown in Figure 1.

Study 1: The Role of Design and Number of Points per Phase

One criticism of the study by Matyas and Greenwood (1990) and voiced against other studies of visual inspection is that they presented only AB (baseline and intervention) design graphs. Therefore, Greenspan and Fisch (1992) generated point-to-point functions to construct AB and ABA (baseline, intervention, and return to baseline) design graphs similar to those in Figure 1 and asked graduate students trained in behavior analysis to identify treatments or trends in 48 graphs. To obtain a systematic replication their results, Greenspan and Fisch employed many of the same values used by Matyas and Greenwood (1990) to generate level shifts, trends, and variability. Level shift was based on effect size. According to Cohen (1988), effect size between two sample means is a ratio of their difference to the common standard deviation, or Δ/σ. An effect size of 1.0 is considered large, producing a power estimate of 84%. The maximum effect size used by Greenspan and Fisch was 2.0, which produced power of 95%.

Greenspan and Fisch (1992) also examined the number of points per phase (5 or 10) to determine whether they affected subjects' responses. According

TABLE 1

Contingent probabilities of responses given the presentation of a particular graph type

		Response				
		A	B	C	D	E
Graph	A	$p(A_r/A_s)$ ...			II	
	B		$p(B_r/B_s)$		II	
Type	C			$p(C_r/C_s)$	II	
	D	I	I	I	$p(D_r/D_s)$	I

A ≡ Level treatment effect only.
B ≡ Trend only.
C ≡ Both treatment effect and trend.
D ≡ Neither treatment effect nor trend.
E ≡ Other systematic change.
I ≡ Type I error (false alarm).
II ≡ Type II error (miss).

to Huitema's (1985) survey, the number of points per phase in most behavioral experiments is frequently small. Matyas and Greenwood (1990) and DeProspero and Cohen (1979) used 10 points per phase. Greenspan and Fisch wanted to assess whether fewer points would increase the Type I error. To evaluate the data, they employed an input-output matrix of contingent probabilities, as shown in Table 1.

Although there were some individual differences, as a group subjects were better able to identify treatment effects only (37%) than trends only (2%), whereas identification of treatment-plus-trend functions fell somewhere in between (17%). As might be expected among trained behavior analysts, when neither treatment effects nor trends were generated, subjects were able to identify 70% of those graphs. Other interesting data were found off the main diagonal of the probability matrix. Type II error rates were high for treatment effects alone and trends alone compared to their respective Type I error rates, but not for treatment-plus-trend graphs. The Type II error rate for those graphs was 3%. Subjects generally identified treatment-plus-trend graphs as treatment only (72%).

Six months later, the study was replicated with a second group. Identification of treatment effects was somewhat lower (23%), trend detection was moderately higher (21%), and identification of treatment-plus-trend graphs was higher still (31%). Consistent with the first analysis, Type II error rates for treatment-plus-trend graphs remained low (6%) and were frequently mistaken as treatment only (60%). Problems in detecting trends confirm results obtained by Matyas and Greenwood (1990), Wampold and Furlong (1981), and Furlong and Wampold (1982) and are consistent with the findings of Cleveland and McGill (1984, 1985), who found that angle discriminations were more difficult to make than either position or length discriminations.

To examine responses to ABA compared to AB design graphs, Greenspan and Fisch (1992) pooled the data for the two groups tested. Subjects were better able to identify level treatment effects alone in ABA designs (37% vs. 26%). However, detection of trends alone or treatment-plus-trend graphs was about the same for AB as for ABA designs. As regards numbers of points per phase, more seemed to be less. Subjects were least successful recognizing treatment effects or trends from ABA graphs with 10 points per phase. Mastery was greater when there were differing numbers of points in adjacent phases (10 and 5, or 5 and 10).

Study 2: The Role of Labeling and Function Placement

Another obstacle in recognizing treatments or trends is that essential information that may have affected judgments of graphs was omitted from earlier studies of visual inspection, as Parsonson and Baer (1992) noted. Fisch and Schneider (1993) thought that one such factor might be the type of response measure depicted. Concurrent with Study 1, they examined the effect of labeling the *y* axis either "Response Rate" or "Proportion of Responses." Fisch and Schneider (1993) also investigated placement of point-to-point functions in the graph: near the top of the figure, in the middle, or near the *x* axis. As in Study 1, Study 2 was replicated with another group of subjects 6 months later. Unlike Study 1, the contingent probabilities computed for the matrices of the two groups in Study 2 showed little difference between them.

Responses from the two groups were pooled and showed once again that level treatment effects were most readily detected (35%), trends were less well recognized (19%), and treatment-plus-trend graphs were interpreted least well (10%). The ability to identify graphs in which neither treatment effects nor trends were present remained high (62%). On the other hand, failure to detect the trend component in treatment-plus-trend graphs was also high (48%).

When Fisch and Schneider (1993) compared subjects' responses to graphs whose *y* axes were labeled "Proportion of Responses" with those labeled "Response Rates," they found that contingent probabilities were nearly identical in all cells of both matrices. On the other hand, function placement had a significant impact on the accurate identification of treatment effects or trends. Specifically, point-to-point functions placed near the top or bottom of the graph elicited a substantially higher proportion of correct responses to treatment effects than those placed in the middle. As a group, functions placed close to the *x* axis generated a higher proportion of correct responses than did those placed elsewhere. Fisch and Schneider (1993) suggested that perhaps the frame of the graph at the top or bottom of the figure provided a visual anchor that enabled the observer to detect changes from baseline more readily than when functions were placed in the middle. Previously, Ottenbacher (1986) attempted to improve detection of treatment effects and trends by displaying celeration lines calculated from the split-middle method of trend estimation. Earlier, Bailey (1984) found that lines of progress improved correct identification of treatment effects and trends, although Crosbie (1987) has argued that the use of trend lines may be problematic when error variance is systematically changed. More recently, Pfadt, Cohen, Sudhalter, Romanczyk, and Wheeler (1992) applied the techniques of statistical process control, developed by Shewhart (1931), to identify outliers by placing upper and lower limits ±3 *SD*s from the moving average to bracket point-to-point functions of a temporal process.

Study 3: The Role of Visual Aids and Experience

Given the results in Study 2 and findings by Bailey (1984), Ottenbacher (1986), and Pfadt et al. (1992), Lee and Fisch (unpublished data) decided to examine the effect of additional guidelines in detecting level treatment effects and trends. They selected 24 graphs from Greenspan and Fisch (1992) along with 24 graphs from Fisch and Schneider (1993), and then drew upper and lower limits (±3 *SD*) on either side of the point-to-point functions in each of the 48 graphs. These, along with the original 48 graphs without upper and lower boundaries, were presented to 5 university faculty members trained as behavior analysts who had many years' experience using visual inspection.

Practiced observers detected a high proportion of level treatment effects (57%) and an extremely high percentage of graphs in which neither treatment effects nor trends were present (91%). However, experienced raters rarely identified trends (1%) or treatment effects-plus-trends (6%); rather, they recognized the treatment effect component only on 63% of the treatment-plus-trend graphs (Lee & Fisch, unpublished data). Where applicable, earlier responses to equivalent graphs from Greenspan and Fisch (1992) were compared with outcomes from Study 3. Seasoned faculty members identified a higher proportion of treatment effects than did graduate students in Study 1 (52% vs. 37%). Faculty members also recognized a higher percentage of graphs in which neither treatment effects nor trends were present (90% vs. 67%). Likewise, responses to equivalent graphs from Fisch and Schneider (1993) compared to those from Study 3 indicated that experienced faculty members identified a higher proportion of treatment effects than did graduate students (52% vs. 32%), and recognized a higher percentage of graphs in which neither treatment effects nor trends were present (90% vs. 57%). These results are in accord with those obtained by Austin and Mawhinney (1996), who found that skilled observers examine graphic data more carefully than novices do.

Responses to graphs with and without guidelines were compared to determine whether visual aids helped observers to interpret functions more accurately. Contingent probability matrices for responses to graphs with guidelines were nearly identical to those without. Thus, experience appears to be more salient than visual guidelines in detecting either treatment effects or the absence of treatment effects and trends. This would explain the results of Bailey (1984), who found that, among graduate students, progress lines produced small increases in percentage of correct responses, whereas Ottenbacher (1986) found that celeration lines were ineffective for many experienced therapists.

Study 4: The Effect of Plotting Trends As an Aid to Visual Inspection

Given the results from Study 3, Fisch and Porto (1994) examined whether the effect of constructing trend lines themselves would influence the way in which observers responded to point-to-point autoregressive functions. To reduce the error rate in detecting trends, these researchers increased the autoregressive coefficent from .30 to .35. They also manipulated the frequency of occurrence of the nonzero autoregressive component, introducing it into one, two, or all three phases of the graphs presented. Subjects were asked first to draw the best fitting line by eye to each of the point-to-point functions in the graph. Then, as in the previous studies, they were asked to circle the response that best described the point-to-point function in the treatment component. To determine reliability, they tested subjects (7 graduate students plus instructor) at the beginning and at the end of a one-semester course on single-subject design.

Best fitting lines drawn by subjects were compared to ordinary least squares regression lines fitted to the point-to-point functions. Two types of errors were identified: (a) a rotational error (R), defined as an angle between the regression line and the eyeball estimate greater than 10°; and (b) a translational error (T), defined as an orthogonal distance between the regression line and the eyeball estimate greater than 0.5 cm. In addition to the usual probability matrices, proportion of correct best fitting line responses as well as proportion of correctly circled responses for test and retest graphs were calculated.

All subjects were better at constructing trend lines than identifying trends. Initially, subjects' correct best fitting line responses averaged 60% (±8%). On retest, all but 1 subject showed improvement. Mean correct line-fitting response was 71% (±7%). Correctly circled responses were much lower, averaging 36% (±5%) initially and 34% (±4%) on retest.

Contingent probability matrices for initial and retest trials were not very different from one another and were very similar to the patterns observed among graduate students in Studies 1 and 2.

SUMMARY AND DISCUSSION

The identification of level treatment effects—a primary consideration for behavior analysts who use visual inspection to interpret graphic data—is the principal focus of trained observers (Fisch & Schneider, 1993; Furlong & Wampold, 1982; Greenspan & Fisch, 1992; Wampold & Furlong, 1981), whatever their level of experience. However, its detection will be affected by moderately noisy data (Greenspan & Fisch, 1992; Matyas & Greenwood, 1990). Therefore, when circumstances limit the degree of experimental control, recognition of treatment effects may be problematic. Practiced observers appear to surmount some obstacles associated with variability by virtue of their own experience. However, they encounter difficulties in detecting trends, whether trends occur by themselves or in combination with treatment effects. In this respect, experienced investigators fare no better than their less well-seasoned counterparts (Fisch & Schneider, 1993; Furlong & Wampold, 1982; Greenspan & Fisch, 1992; Wampold & Furlong, 1981). Spence (1990) and Parsonson and Baer (1992) have conjectured that the problem of trend recognition is a psychophysical matter. If so, the psychophysical parameters of trend detection need to be examined more systematically, as Cleveland and McGill (1987) and McEwan (1994) have attempted recently.

For less experienced observers, guidelines such as the upper and lower limits employed in statistical process control may improve detection of treatment effects or trends. However, there are computational problems when using statistical process control with autoregressive processes (Wetherill & Brown, 1991). Trend lines may be useful as heuristic devices for less experienced observers. But trend line construction does not appear to facilitate trend detection. Where appropriate, it may be more valuable to graph point-to-point functions as near to the x axis as possible in order to exploit the framing features provided by the horizontal and vertical axes.

Parsonson and Baer (1992) have argued that graphs used in earlier studies were not evaluated under the normal conditions of research, in that contextual information was absent (e.g., information regarding the independent and dependent variables). However, Fisch and Schneider (1993) show that at least one of those contextual factors, the type of dependent variable employed, had no effect on detection of trend or treatment effect. That is not to say that contextual information per se is useless. One would hope that the factors cited by Parsonson and Baer (1992) could be investigated systematically to determine which were effective.

Type II errors, which continue as a hallmark of visual inspection for behavioral analysis, remain high for all observers. However, contrary to remarks made by Sidman (1960), behavior analysts who rely solely on visual inspection will not be sensitive to small but conceivably important differences, especially those that may arise from attempting to shape onerous daily living skills or complex behaviors in humans. Applied behavior analysts would thus be unable to capitalize on the modest but meaningful successes that may have transpired.

To determine changes between phases, one could estimate effect size based on obtained differences among baseline, intervention, and return-to-baseline phases. Cohen (1988) has written extensively and imploringly on the use of effect size and power analysis in conjunction with significance testing, although effect size may be more appropriate in computing level shifts than autocorrelated trends. Others (e.g., Hahn & Meeker, 1991) have argued for the use of confidence intervals to quantify uncertainty associated with point estimates. There is some difference of opinion regarding the utility of statistical intervals on data that evolve from temporal processes as opposed to well-defined populations, but practictioners of statistical process control have long and successfully argued for their implementation (Deming, 1953).

Despite its air of simplicity, the visual inspection of graphic data is a more complex and subtle process than thought originally. If behavior analysts are to optimize the utility of their tools for data analysis, they will have to consider methods in addition to those currently deemed useful. One such technique may be to employ observers' verbal reports as data to modify the observers' performance on visual inspection tasks (Austin & Mawhinney, 1996; Ericsson & Simon, 1980). Another is the use of feedback, suggested by Grote and Baer (1996). A third may be the use of signal-detection analysis to identify sources of error, as noted by Karp and Fisch (1996).

One is also tempted to suggest, as others have over the past several decades, that perhaps the time has come to reappraise statistical procedures for use in behavior analysis, particularly those that have been developed recently (e.g., Haccou & Meelis, 1992; Magnusson, Bergman, Rudinger, & Törestad, 1991). Problems associated with the application of frequentist statistics and conventional time series analysis have been addressed previously in the literature, and arguments for and against significance testing have reappeared recently (e.g., Shrout, 1997). Formal arguments notwithstanding, the case against the use of statistics by behavior analysts can be reduced to one of effect magnitude and whether a statistically significant difference translates into a clinically or experimentally meaningful one (Baer, 1977; Kazdin, 1978). This would have greater validity if identification of level treatment effects were not the primary but sole objective of visual inspection. However, trends, which are also an important characteristic of behavioral data, have not been easily detected by visual inspection and need to be ascertained more readily.

One solution to trend detection may be to analyze temporal responses by means of time-structured Markov chains (Haccou & Meelis, 1992). A Markov chain is a sequence of two-dimensional arrays (matrices) that contain the probabilities of changing from one set of conditions to another, from one time period to the next. It is a model often used to evaluate behavior that occurs in naturalistic settings, when variability and duration of specific behaviors make a deterministic model infeasible. The "state" of the system at a given moment, along with the transition probability matrix, contain all the information necessary about the system to calculate probabilities of future changes. Another solution may be to estimate serial correlation using both visual inspection and statistical techniques, as described by Stigler (1986). As with similar issues, this question can be answered empirically. One should also bear in mind the remarks made by Skinner (1938) and alluded to earlier, that perhaps the time has come to reevaluate the current state of statistical and nonstatistical conceptions of science and whether they are truly at variance with one another.

REFERENCES

Austin, J., & Mawhinney, T. (1996, May). *Relations among data analysts' discriminations between baseline and intervention streams, methods of data presentation, and analyses.* Paper presented at the 22nd annual convention of the Association for Behavior Analysis, San Francisco.

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.

Bakan, D. (1954). A generalization of Sidman's results on group and individual functions, and a criterion. *Psychological Bulletin, 51,* 63–64.

Busk, P. L., & Marascuilo, L. A. (1988). Autocorrelation in single-subject research: A counterargument to the myth of no autocorrelation. *Behavioral Assessment, 10,* 229–242.

Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association, 79,* 531–554.

Cleveland, W. S., & McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science, 229,* 828–833.

Cleveland, W. S., & McGill, R. (1987). Graphical perception: The visual decoding of quantitative information on graphical displays of data. *Journal of the Royal Statistical Society, A, 150,* 192–229.

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

Crosbie, J. (1987). The inability of the binomial test to control Type I erro with single-subject data. *Behavioral Assessment, 9,* 141–150.

Crosbie, J. (1989). The inappropriateness of the *C* statistic for assessing stability or treatment effects with single-subject data. *Behavioral Assessment, 11,* 315–325.

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

Deming, W. E. (1953). On the distinction between enumerative and analytic surveys. *Journal of the American Statistics Association, 48,* 244–255.

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

Ericsson, A. K., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87,* 215–281.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin, 53,* 134–140.

Fisch, G. S., & Porto, A. F. (1994). Visual inspection of data: Does the eyeball fit the trend? In B. E. Rogowitz & J. P. Allebach (Eds.), *Human vision, visual processing, and digital display V* (pp. 268–276). Bellingham, WA: SPIE Volume 2179.

Fisch, G. S., & Schneider, R. (1993). Visual inspection of data: Placement of point-to-point graphs affects discrimination of trends and treatment effects. *American Statistical Association Proceedings: Section on Statistical Graphics,* 55–58.

Furlong, M. J., & Wampold, B. E. (1982). Intervention effects and relative variation as dimensions in experts' use of visual inference. *Journal of Applied Behavior Analysis, 15,* 415–421.

Gentile, J. R., Roden, A. H., & Klein, R. D. (1972). An analysis-of-variance model for the intrasubject replication design. *Journal of Applied Behavior Analysis, 5,* 193–198.

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

Greenspan, P., & Fisch, G. S. (1992). Visual inspection of data: A statistical analysis of behavior. *American Statistical Association Proceedings: Section on Statistical Graphics,* 79–82.

Greenwood, K. M., & Matyas, T. A. (1990). Problems with the application of interrupted time series analysis for brief single-subject data. *Behavioral Assessment, 12,* 355–370.

Grote, I., & Baer, D. M. (1996, May). *Visual analysis of graphic data by scientists-in-training.* Paper presented at the 22nd annual convention of the Association for Behavior Analysis, San Francisco.

Haccou, P., & Meelis, E. (1992). *Statistical analysis of behavioural data: An approach based on time-structured models.* Oxford, England: Oxford University Press.

Hahn, G. J., & Meeker, W. Q. (1991). *Statistical intervals: A guide for practitioners.* New York: Wiley.

Harrop, J. W., & Velicer, W. F. (1985). A comparison of alternative approaches to the analysis of interrupted times-series. *Multivariate Behavioral Research, 20,* 27–44.

Hartmann, D. P. (1977). Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis, 10,* 103–116.

Hartmann, D. P., Gottmann, J. M., Jones, R. R., Gardner, W., Kazdin, A. E., & Vaught, R. S. (1980). Interrupted time-series analysis and its application to behavioral data. *Journal of Applied Behavior Analysis, 13,* 543–559.

Huitema, B. E. (1985). Autocorrelation in applied behavior analysis: A myth. *Behavioral Assessment,* 109–120.

Huitema, B. E. (1986). Autocorrelation in behavioral research: Wherefore art thou? In A. Poling & R. W. Fuqua (Eds.), *Research methods in applied behavior analysis: Issues and advances* (pp. 187–208). New York: Plenum Press.

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

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

Karp, H. J., & Fisch, G. S. (1996, May). *Signal detection analysis of responses to graphic data produced by auto-correlation and least squares regression equations.* Paper presented at the

22nd annual convention of the Association for Behavior Analysis, San Francisco.

Kazdin, A. E. (1977). Artifact, bias, and complexity of assessment: The ABCs of reliability. *Journal of Applied Behavior Analysis, 10,* 141–150.

Kazdin, A. E. (1978). Statistical analysis for single-case experimental designs. In M. Hersen & D. H. Barlow (Eds.), *Single case experimental designs: Strategies for studying behavioral change* (pp. 265–316). New York: Pergamon Press.

Kelly, M. B. (1977). A review of the observational data-collection and reliability procedures reported in the *Journal of Applied Behavior Analysis. Journal of Applied Behavior Analysis, 10,* 97–101.

Keselman, H. J., & Leventhal, L. (1974). Concerning the statistical procedures enumerated by Gentile et al.: Another perspective. *Journal of Applied Behavior Analysis, 7,* 643–645.

Kratochwill, T., Alden, K., Demuth, D., Dawson, D., Panicucci, C., Arntson, P., McMurray, N., Hempstead, J., & Levin, J. (1974). A further consideration in the application of an analysis-of-variance model for the intrasubject replication design. *Journal of Applied Behavior Analysis, 7,* 629–633.

Kratochwill, T. R., & Wetzel, R. J. (1977). Observer agreement, credibility, and judgment: Some considerations in presenting observer agreement data. *Journal of Applied Behavior Analysis, 10,* 133–139.

Lewandowsky, S., & Spence, I. (1989). Discriminating strata in scatterplots. *Journal of the American Statistical Association, 84,* 682–688.

Magnusson, D., Bergman, L. R., Rudinger, G., & Törestad, B. (1991). *Problems and methods in longitudinal research: Stability and change.* Cambridge, England: Cambridge 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. (1991). Problems in the estimation of autocorrelation in brief time series and some implications for behavioral data. *Behavioral Assessment, 13,* 137–157.

McEwan, J. S. A. (1994). *Optimising line graph aspect ratio.* Unpublished doctoral dissertation, The University of Waikato, New Zealand.

Michael, J. (1974). Statistical inference for individual organism research: Mixed blessing or curse? *Journal of Applied Behavior Analysis, 7,* 647–653.

Ottenbacher, K. J. (1986). Reliability and accuracy of visually analyzing graphed data from single-subject designs. *American Journal of Occupational Therapy, 40,* 464–469.

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. 15–41). Hillsdale, NJ: Erlbaum.

Pfadt, A., Cohen, I. L., Sudhalter, V., Romanczyk, R. G., & Wheeler, D. J. (1992). Applying statistical process control to clinical data: An illustration. *Journal of Applied Behavior Analysis, 25,* 551–560.

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

Shewhart, W. A. (1931). *Economic control of quality of manufactured product.* Princeton, NJ: Reinhold Co.

Shrout, P. E. (1997). Should significance testing be banned? Introduction to a special section exploring the pros and cons. *Psychological Science, 8,* 1–2.

Sidman, M. (1952). A note on functional relations obtained from group data. *Psychological Bulletin, 49,* 263–269.

Sidman, M. (1960). *Tactics of scientific research: Evaluation of experimental data in psychology.* New York: Basic Books.

Skinner, B. F. (1938). *The behavior of organisms.* New York: Appleton-Century-Crofts.

Spence, I. (1990). Visual psychophysics of simple graphical elements. *Journal of Experimental Psychology: Human Perception and Performance, 16,* 683–692.

Spence, I., & Lewandowsky, S. (1990). Graphical perception. In J. Fox & J. S. Long (Eds.), *Modern methods of data analysis* (pp. 13–57). Newbury Park, CA: Sage.

Stigler, S. M. (1986). Estimating serial correlation by visual inspection of diagnostic plots. *The American Statistician, 40,* 111–116.

Thoresen, C. E., & Elashoff, J. D. (1974). An analysis-of-variance model for intrasubject replication design: Some additional comments. *Journal of Applied Behavior Analysis, 7,* 639–641.

Toothaker, L. E., Banz, M., Noble, C., Camp, J., & Davis, D. (1983). N = 1 designs: The failure of ANOVA-based tests. *Journal of Educational Statistics, 8,* 289–309.

Velicer, W. F., & Harrop, J. (1983). The reliability and accuracy of time series model identification. *Evaluation Review, 7,* 551–560.

Wampold, B., & Furlong, M. (1981). The heuristics of visual inference. *Behavioral Assessment, 3,* 79–92.

Wetherill, G. B., & Brown, D. W. (1991). *Statistical process control: Theory and practice.* London: Chapman and Hall.

Yeltman, A. R., Wildman, B. G., & Erickson, M. T. (1977). A probability-based formula for calculating interobserver agreement. *Journal of Applied Behavior Analysis,* 127–131.