VISUAL ANALYSIS OF SINGLE-CASE TIME SERIES: EFFECTS OF VARIABILITY, SERIAL DEPENDENCE, AND MAGNITUDE OF INTERVENTION EFFECTS

THOMAS A. MATYAS AND KENNETH M. GREENWOOD

LA TROBE UNIVERSITY, VICTORIA, AUSTRALIA

Visual analysis is the dominant method of analysis for single-case time series. The literature assumes that visual analysts will be conservative judges. We show that previous research into visual analysis has not adequately examined false alarm and miss rates or the effect of serial dependence. In order to measure false alarm and miss rates while varying serial dependence, amount of random variability, and effect size, 37 students undertaking a postgraduate course in single-case design and analysis were required to assess the presence of an intervention effect in each of 27 AB charts constructed using a first-order autoregressive model. Three levels of effect size and three levels of variability, representative of values found in published charts, were combined with autocorrelation coefficients of 0, 0.3 and 0.6 in a factorial design. False alarm rates were surprisingly high (16% to 84%). Positive autocorrelation and increased random variation both significantly increased the false alarm rates and interacted in a nonlinear fashion. Miss rates were relatively low (0% to 22%) and were not significantly affected by the design parameters. Thus, visual analysts were not conservative, and serial dependence did influence judgment.

DESCRIPTORS: visual inference, data analysis, single-subject design

Despite the acknowledged dominance of visual analysis in single-case methodology (Kazdin, 1982) research on the performance of visual analysts is relatively sparse (DeProspero & Cohen, 1979; Furlong & Wampold, 1982; Jones, Weinrott, & Vaught, 1978; Ottenbacher, 1986; Wampold & Furlong, 1981). The published data have indicated significant problems such as poor interjudge reliability (DeProspero & Cohen, 1979; Jones et al., 1978). Some of these results (Jones et al., 1978) have been attacked for poor methodology (Huitema, 1985). Notwithstanding these problems, visual analysis has received continued advocacy (Parsonson & Baer, 1986).

The lack of evidence does not seem to have deterred claims about the likely performance of visual analysts (Kazdin, 1982; Parsonson & Baer, 1978, 1986). In particular, it is claimed that visual analysts are more likely to commit Type II errors "than those relying on statistical analyses" (Kazdin, 1982, p. 242). Similarly, Parsonson and Baer venture that: "If changes in graphed data are to be seen as such, they need to be relatively large—so large that the visual analysis of data tends to be less sensitive than statistical analysis of the same data. . . . Insensitivity ought to generate more conservative judgments that behavior has changed in correlation with experimental variables" (Parsonson & Baer, 1986, p. 158).

This however is a shallow deduction given the relationship between Type I (false alarm) and Type II (miss) errors. If visual analysts are insensitive as claimed (and this remains to be empirically demonstrated), this does not guarantee that they will rarely produce false alarms. Visual analysts may simply be more noisy detectors. That is, they may both miss and produce false alarms at a high rate. It requires a further assumption to argue that the human judge will be a detector with desired low false alarm rates and low sensitivity. This assumption is that the visual analyst will give to the control of Type I errors the same high priority that has been given to it in statistical decision theory. However, the literature has not adequately addressed the question of false alarm and miss rates in visual analysis, although a number of papers have investigated the performance of analysts (DeProspero & Cohen, 1979; Furlong & Wampold, 1982; Jones

Requests for reprints should be sent to T. A. Matyas, Department of Behavioural Health Sciences, Lincoln School of Health Sciences, La Trobe University, 625 Swanston Street, Carlton, Victoria, 3053 Australia.

et al., 1978; Ottenbacher, 1986; Wampold & Furlong, 1981).

Jones et al. (1978) investigated interobserver agreement rates and agreement between the visual and statistical interrupted time series analysis (ITSA) of published case data. They concluded that experienced judges had poor agreement rates. If their statistical analyses are taken as a yardstick, their results imply a false alarm rate of 33% and a miss rate of 48% for experienced judges, figures that should create significant concern. Unfortunately, the analysis of Jones et al. was flawed in some respects (Huitema, 1986a). Most importantly, however, the method of comparing the performance of visual analysts against statistical analysis cannot address the issue of misses and false alarms unambiguously, particularly in the absence of a power analysis. We cannot deduce that the visual analyst is wrong if the human judge declared an effect when the statistical analysis found no significant effect. An effect might have existed and the statistical analysis possessed insufficient power (sensitivity) to detect it, whereas the human judge, with unknown operating characteristics, may have detected the effect correctly. Low power was extremely likely in the results of Jones et al. because they were not only operating within the limitations of brief phases that are often imposed by case data but in addition they deliberately selected cases with "small number of data points within phases" (Jones et al., 1978, p. 278). A further problem with the study of Jones et al. is that they deliberately biased the selection of the charts to obtain, in addition to cases with brief phases, "nonobvious" experimental results, "graphs where serial dependency might be evidenced by possible non-zero trend" and "excluded large effect experiments" (p. 278). Thus, the implied estimates of false alarm and miss rates that might be deduced from their data cannot be trusted.

DeProspero and Cohen (1979) adopted the strategy of constructing ABAB charts in which they introduced effects. These were submitted to analysis by a large sample of reviewers of behavioral journals. However their results do not allow examination of the false alarm rate, because all graphs had introduced some degree of interphase differences. Further, the judges were required to rate on a 0 to 100 scale the degree of experimental control shown rather than to make a forced-choice decision; this precluded conclusions about miss rates.

Ottenbacher (1986) exposed 46 occupational therapists to five AB panels. It is not clear which charts contained an effect and which did not, thereby precluding the analysis of false alarm and miss rates. He did attempt to analyze the Type I error rate by comparing analysts' decisions with statistical analyses using White's (1974) suggestion for employing the binomial distribution on the intraphase celeration lines. Ottenbacher's approach thus suffers from the limitation discussed above in connection to Jones et al. (1978): Statistical analysis is not an acceptable yardstick for identification of "no effect" when the power is likely to be very low and no power analysis is even attempted. Ottenbacher's data (n = 8 per phase) were very likely a low power case. In any case, White's suggested method of analysis is flawed (Crosbie, 1987).

Wampold and Furlong (1981) asked graduate students to classify AB charts into groups according to the type of effects perceived. Furlong and Wampold (1982) extended this investigation to a sample of expert analysts (10 *JABA* reviewers). Unfortunately, these studies did not include no-effect charts and thus could not address the false alarm issue. However, neither do they report the miss rate, preferring to concentrate on other aspects of judge performance.

Thus, none of the empirical studies conducted to date have adequately addressed the question of false alarm and miss rates in even simple designs such as AB panels. Therefore, one aim of the present study was to examine these directly by requiring forced-choice decisions in AB panels with and without known effects. The ability to make decisions about the basic AB panel may be more fundamental than is generally acknowledged. Although the AB design is one of the weakest case designs, the AB panel represents the building block of more complex decisions in ABAB, multiple baseline, changing criterion, and other more sophisticated designs.

Another issue that has concerned visual analysis is the possibility that serial dependence in the data may alter the accuracy of the analyst. Jones et al. (1978) concluded that visual analysis was adversely affected by increased serial dependence. However, their analysis of autocorrelation was incorrect (Huitema, 1986a). Hence their conclusion that serial dependence affects rater reliability and the degree of agreement with statistical ITSA becomes equivocal. DeProspero and Cohen's (1979) results are consistent with the notion that serial dependence matters, but the conclusion is very indirect. They manipulated the trend in data by tilting the baseline 30° from horizontal. Although they did not give calculations of autocorrelation, the tilting maneuver would have increased the amount of serial dependence in the data. They reported that the absence of trend produced higher subjective confidence of "experimental control." Because they did not calculate autocorrelation, we cannot assess its contribution in baselines that were not tilted. Further, tilting baselines by 30° was an arbitrary way of introducing trend. They offered no evidence that a 30° tilt would introduce serial dependence typical of behavioral data. Ottenbacher (1986) reported that there was only a weak relationship between serial dependence in visual charts and observer disagreement. His analysis was flawed, however, and a reanalysis of his data indicated that a strong relationship existed between serial dependence in the baseline and interjudge disagreement (Matyas & Greenwood, in press).

In summary, the empirical literature on visual analysis to date has failed to examine the adequacy of the process to the extent implicitly demanded by its advocacy as a fundamental method for case management and analysis (Barlow, Hayes, & Nelson, 1984). Although the issue of interjudge reliability has been repeatedly addressed, the basic questions of false alarm and miss rates have not been adequately investigated. The existing studies of visual analysis appear also to have been concerned directly, or indirectly, with the effect of serial dependence. However, methodological problems limit the conclusions possible from these studies. Therefore the present study aimed to quantify false alarm and miss rates, which are the fundamentals of decision making. Serial dependence in the time series was also systematically varied with reference to two surveys of the degree of autocorrelation in published data (Huitema, 1985; Matyas & Greenwood, 1985, 1990).

METHOD

Subjects

The sample comprised 37 graduate students from two groups ($n = 18$, $n = 19$) undertaking one-term courses in single-case design and analysis. The majority ($n = 25$) had obtained a bachelor's qualification with a major in psychology prior to enrollment in the course. The remainder comprised practicing health professionals (6 occupational therapists, 2 physiotherapists, 2 orthoptists, 1 neurologist, 1 podiatrist) who had typically completed only a 2-year minor in psychology. None of the subjects were experienced users of single-case designs. The investigation reported below was conducted after these students were exposed to a series of lectures on single-case design. The nature of level, trend, and other intervention effects had been discussed, as had been difficulties introduced by high variability and preexisting trend. The empirical literature on visual analysis was not reviewed until after the data collection session. Assigned readings for the course, up to the point of data collection, were from the textbooks by Hersen and Barlow (1976), Kratochwill (1978), and Kazdin (1982).

Materials

Twenty-seven AB (A = baseline, B = intervention) panels were constructed using the first order autoregressive model: $y_t = ay_{t-1} + b + d + e$, where y_t was the value at time t, y_{t-1} was the value at time $t - 1$, a was the autoregression coefficient, b was the preintervention initial level, d was the intervention effect, and e was a normally distributed random variable with a mean of 0 and standard deviation described below. Each phase comprised 10 data points, which seemed a reasonable value in the light of Huitema's (1985) survey. It should be noted that we have previously investigated the

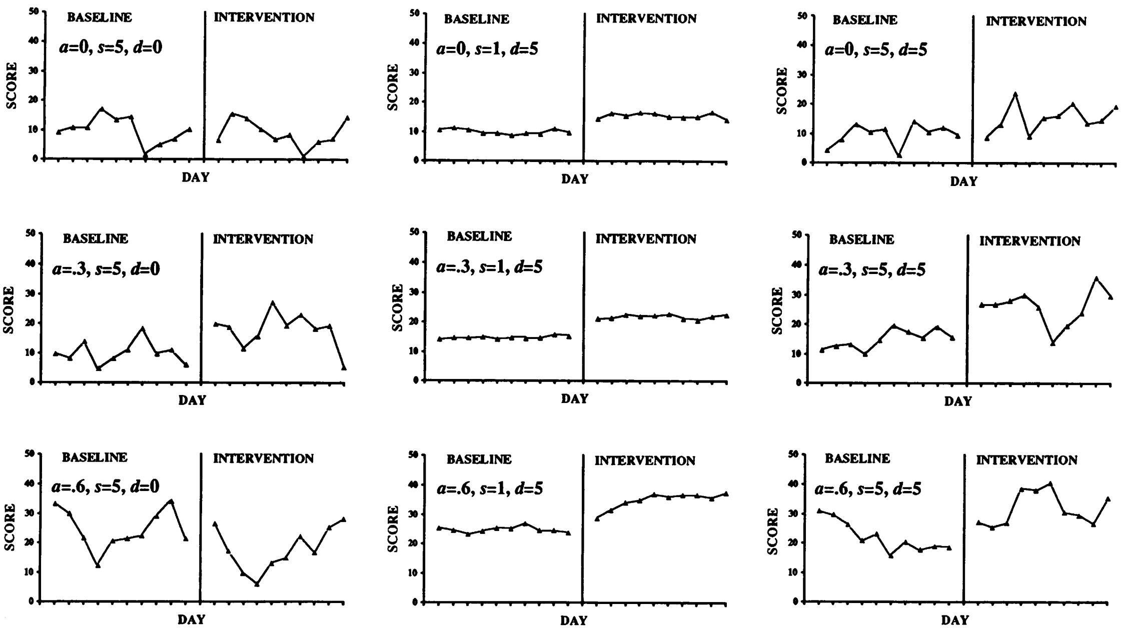

Figure 1. Nine of the AB panels used as stimuli. Information about the statistical properties (*a* = amount of serial dependence, *s* = random variability, *d* = magnitude of intervention effect) has been superimposed on each graph and was not presented to subjects. Subjects were instructed that the ordinate represented the client's response score, but the precise response was not described.

viability of autoregressive models for serial dependence in baseline data and found the first order model reasonable (within the power limits) for the vast majority (91%) (Matyas & Greenwood, 1985). Figure 1 illustrates some of the charts yielded by this method.

Three factors, each with three levels, were varied to obtain charts for a completely crossed factorial design. One factor was the effect of intervention that involved null and two magnitudes of treatment effect (i.e., $d = 0$, $d = 5$, $d = 10$). A second factor was the amount of random error chosen ($s = 1$, $s = 3$, $s = 5$). The third factor was the degree of serial dependence in the data. Our survey of 182 baselines published in *JABA* from 1977 to 1983 (Matyas & Greenwood, 1985), as well as that by Huitema (1985) covering a complementary period, suggested a range of first-order autocorrelation from -0.68 to 0.75, with a mean around zero. When corrected for a recently verified bias in the autocorrelation estimation procedure (Matyas & Greenwood, 1990), these data suggest a true range from -0.80 to 0.90. Because previous investigations have been primarily concerned with the effects of positive autocorrelation but have some methodological limitations, we chose to investigate three levels of true autocorrelation: $a = 0.0$, $a = 0.3$, and $a = 0.6$. These are equivalent to estimated autocorrelation of -0.1, 0, and 0.2 when $n = 10$. According to both Huitema (1985) and Matyas and Greenwood (1990), values like these are common in published time series. The effect sizes selected are best understood as the d/s ratio following Cohen (1977). Thus, the standardized effect sizes employed ranged from 1 to 10. These would be described as large to very large standardized effects by Cohen (1977), who regarded standardized effect sizes of 1 or more as representing large effects in the social/behavioral sciences. To provide an additional frame of reference for our values, time series from our *JABA* survey (Matyas & Greenwood, 1985) were analyzed according to Gottman's simplified ITSA method (Gottman, 1981) using the Gottman–Williams software suite (Williams & Gottman, 1982). A full description of this and related analyses is beyond the scope of this paper. However, the median d/s ratio obtained from 100 AB panels with $n \geq 10$ was 9.2, the 25th percentile was 4.9 and the 75th was 17.1. Thus, our standardized effect size of 1 appears to be below the 25th percentile of effect sizes published in *JABA*, and our standardized effect size of 10 appears to be just above the median effect size published in *JABA*.

Charts were produced on an Apple Macintosh Plus® computer using Microsoft Chart® software and were labeled as depicted in Figure 1, except that Figure 1 provides values for a, s, and d. Overhead transparencies were than obtained in A4 format for group presentation.

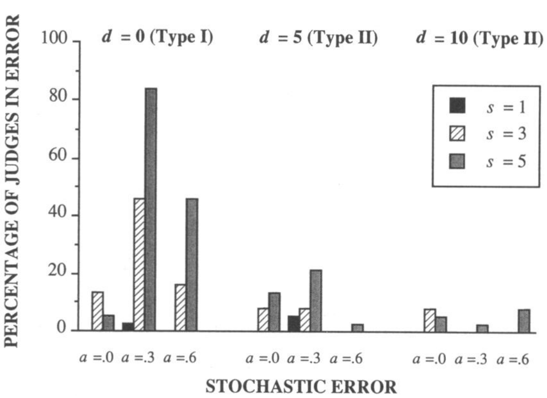

Figure 2. Type I and II error rates for each of the 27 charts as a function of the amount of serial dependence (a), random variability (s), and magnitude of intervention effect (d).

Procedure

Subjects were tested in two separate groups ($n = 18$, $n = 19$) in single 1-hr sessions. Subjects were initially instructed how to respond on a standard computer card for recording answers to multiple choice questions. The alternatives were defined as follows: A = no intervention effect; B = a level change; C = a trend change; D = combined level and trend change; and, E = other type of systematic change during intervention. A brief review of treatment effect types was conducted prior to data collection, with ideal examples from Glass, Willson, and Gottman (1975) consistent with previous lecture material on this subject.

The 27 charts were presented in a predetermined, randomized sequence. Charts were presented in a small room on a large screen. Projection clarity from all seats was established by the experimenters prior to the session. Each chart was presented for 1 min. Prior pilot data had suggested that 20 to 40 s be allowed for ample response times, depending on the individual and the chart. Thus, the test was not presented as a speed test and subjects were told to ask for additional time if required. None did. All subjects viewed and responded to all 27 charts.

RESULTS

Responses for each subject on each case chart were scored dichotomously according to whether they indicated a conclusion of effect (Alternatives B to E) or no effect (Alternative A) because the major interest was in false alarm and miss rates. These responses were then analyzed by a series of planned comparisons, employing the extension to Cochran's Q detailed by Marascuilo and McSweeney (1977). This method suited the analytic problem, not only because it deals legitimately with dichotomous data generated by a within-subject design but also because, in the planned comparison version, it permitted the phrasing of the main effect and interaction questions that arose naturally from the factorial design.

Figure 2 presents the percentage of judges who erred in each of the 27 cases comprising the experimental design: that is, three levels of autocorrelation (a = 0.0, 0.3, or 0.6) × three levels of random variability (s = 1, 3, or 5) × three levels of effect size (d = 0, 1, or 10). The figure shows that high false alarm rates occurred when variability and serial dependence increased. Error rates ranged from 16% to 84% when $s > 1$ and $a > 0$. These high false alarm rates contrasted with the relatively low miss rates, which ranged from 0% to no more than 22%. Most of the miss rates were below 10%.

The planned comparisons confirmed that the apparent general difference between miss and false alarm rates was statistically significant, $z = 2.499$, $p < .02$. The comparisons confirmed further that this difference became more accentuated as the random variation increased $z = 4.79$, $p < .001$, and as the degree of serial dependence increased, $z = 4.68, p < .001$.

Because the comparison designed to examine the three-way interaction between effect condition, degree of random variation, and degree of autocorrelation was significant, comparisons were constructed to investigate the two-way interaction between degree of random variation and degree of autocorrelation independently for false alarm and miss data, analogously to recommendations typical for factorial analysis of variance (Keppel, 1982). This orthogonal partitioning revealed that the false alarm rate showed a significant interaction between degree of variability and degree of serial dependence, $z = 2.657, p < .01$, which was also significant under the more stringent Dunn–Bonferroni criterion (Marascuilo & McSweeney, 1977). This analysis confirms the impression conveyed by Figure 2 that false alarm rates increased at $s = 3$ and more so at $s = 5$, but that these increases occurred more markedly in the presence of positive autocorrelation, particularly at $a = 0.3$. Variations in miss rates were generally of lower magnitude, and the corresponding two-way interaction comparison for these data was not significant.

In view of the significant interaction between degree of autocorrelation and degree of random variation, the effects of autocorrelation of false alarm rates were examined with six orthogonal comparisons. At $s = 1$, the averaged false alarm rates for $a = 0.3$ and $a = 0.6$ did not differ significantly from that obtained with $a = 0$. However the corresponding differences were significant at $s = 3$, $z = 2.052, p < .05$, and $s = 5$, $z = 5.998, p < .001$. Significant pairwise differences were then obtained between $a = 0.3$ and $a = 0.6$ at $s = 3$, $z = 3.005, p < .01$, and $s = 5$, $z = 3.824, p < .001$, but not at $s = 1$. All comparisons except that between the average false alarm rates at $a = 0.3$ and $a = 0.6$ against $a = 0$ at $s = 3$ were also significant under the more conservative Dunn–Bonferroni criterion for a six-comparison family. Curiously, as the figure and the comparisons both indicate, the false alarm rate was affected nonlinearly by positive autocorrelation, with stronger effects at $a = 0.3$ than at $a = 0.6$.

The simple main effects of degree of random variation on false alarm rates were also pursued, given the significant interaction between degree of autocorrelation and degree of random variation on false alarm rates, in the manner recommended for analysis of variance (see Keppel, 1982). Six pairwise contrasts were employed. At $a = 0$ the difference between false alarm rates with $s = 1$ was not significantly different from that of $s = 3$ or $s = 5$. However, at $a = 0.3$ the false alarm rate with $s = 1$ was significantly lower than with $s = 3$, $z = 4.372, p < .001$, and with $s = 5$, $z = 5.467, p < .001$. At $a = 0.6$ the false alarm rate with $s = 1$ was significantly lower than with $s = 5$, $z = 4.646, p < .001$, but not significantly lower than the false alarm with $s = 3$. All three significant comparisons were also significant under the Dunn–Bonferroni criterion. In general, the investigative comparisons confirmed that the interaction between degree of autocorrelation and random error was due to increases in false alarm rates obtained with higher values of random variation at $a = 0.3$ and $a = 0.6$, but not at $a = 0$, when false alarm rates were generally low.

The above analysis examined trends as a function of the experimental conditions. Individual differences in false alarm rates and miss rates were of interest in order to assess whether subjects tended to vary in their response bias. Thus, false alarms were calculated for each subject over the sample of 9 "no effect" cases. Miss rates were calculated for each subject over the 9 cases with an intervention effect of 5 units and also over the 9 cases with an intervention effect of 10 units. Pearson's correlation coefficients were calculated between false alarm and miss rates. Most interestingly, a significant inverse relationship was found between the false alarm rates and the miss rates at $d = 5$, $r = -0.43, p < .005$, and at $d = 10$, $r = -0.36, p < .02$.

DISCUSSION

The results indicate that our sample of graduate students (who were relatively inexperienced in the assessment of single-case time series), when exposed to AB panels with stochastic properties that are representative of published cases, showed generally high false alarm rates and relatively low miss rates. False alarm rates tended to increase when random variation increased, but only if positive true lag 1 autocorrelation was also present in the data. Similarly, positive autocorrelation tended to increase false alarm rates, but not if the random variation was very low. There was some evidence, however, that this increase may not progress linearly as a function of increasing positive autocorrelation. In general, positive autocorrelation and random variation tended to increase false alarm rates under a mutually potentiating interaction.

That the high false alarm rate is merely a function of the relative inexperience of the participants is possible but seems unlikely. It should be noted that our sample consisted of graduates in psychology undertaking a postgraduate course in health psychology and also included other health professionals. The sample had also received some training prior to testing. Most importantly, studies that have examined experienced judges (DeProspero & Cohen, 1979; Furlong & Wampold, 1982; Jones et al., 1978), although not able to address the false alarm issue satisfactorily, have found indications of inaccurate or unreliable decision making. Studies that have compared the performance of analysts with differing levels of experience (Furlong & Wampold, 1982; Knapp, 1983) have found no major differences in their performance indicators. In any case, the performance of relatively inexperienced judges is of significant interest, given that this may be the skill level currently representative of clinicians. Until our study is replicated in other samples and extended with other designs, it is difficult to judge fully the degree of concern that should be directed to the functioning of the wider population of clinicians who are being encouraged by the literature to adopt experimental single-case methodology. It seems reasonable at least to propose that the assumption that visual judges will be necessarily conservative does not hold for beginning practitioners employing AB designs under conditions of moderate to high variability and serial dependence.

The high false alarm rate obtained confirms the criticism suggested earlier of the position taken by some authors (e.g., Parsonson & Baer, 1986), who seem to have prematurely inferred the dominance of conservatism in human judgment of case charts. Thus, although there may be comfort in the finding that effects are not missed, as anticipated (for large effects) by Parsonson and Baer (1986), the much more serious problem of high false alarm rates, not envisaged by Parsonson and Baer, may be frustrating the valid development of single-case methodology for clinical practice. This seems to be a particularly serious issue because a fundamental argument for introducing the rigors of single case experimental method to clinical practice is to aid valid decision making. The lack of conservatism is not entirely surprising in light of other extensive literature that demonstrates several biases in human decision making (e.g., Hogarth, 1980; Slovic, Fischhoff, & Lichtenstein, 1977). The catalogue of biases includes underestimation of error and undue confidence in small samples (Kahneman & Tversky, 1972; Tversky & Kahneman, 1971), as well as unduly narrow estimation of confidence intervals (Lichtenstein, Fischhoff, & Phillips, 1977) and illusory correlation (Chapman & Chapman, 1969). This lack of conservatism, which has now also been illustrated in the context of single-case time series, indicates the potential for significant practical problems in routine clinical methodology if active clinicians have performances comparable to our sample.

Our findings contribute towards the resolution of the problem about the true effects of serial dependence on visual judgment. This problem resulted from the inadequacies in Jones et al.'s (1978) methodology, detailed in the introduction, its replication and other confounding effects that occurred in Ottenbacher's (1986) study, and the arbitrary choices of tilt in DeProspero and Cohen's (1979) studies. In our sample, positive autocorrelation in amounts that do occur in published data (Matyas & Greenwood, 1985, 1990) increased the false alarm rate, particularly when random variability was larger. These results confirm our reanalysis of Ottenbacher's (1986) study (Matyas & Greenwood, in press).

It is not immediately apparent why false alarm rates would be increased most markedly by the conjunction of positive autoregression and larger random variation. A time series that was purely random, even one with large random variation, is unlikely to demonstrate sustained drifts. Such a time series may present some difficulties in recognizing small effects but is unlikely to have sustained change suggestive of an effect when none is present. Indeed, false alarm rates at $a = 0$ were only 13.5% at $s = 3$ and only 5.4% at $s = 5$. A positive autoregressive process, however, introduces some "inertia" into the time series, such that when a large random component occurs its effects will persist, in a decaying form, through the autoregressive coefficient. Consequently, large random variance in conjunction with positive autocorrelation is more likely to create time series with a larger, apparently systematic trend or change of trend. Only relatively long baselines are likely to permit perception of the true nature of the effect: that of randomly timed but sustained change, which is both positive and negative and of variable duration. Over a relatively short phase, it is more likely than in the case of the purely random time series that the autoregressive stochastic trend will look deterministic. This raises the possibility of several effect-like appearances, such as reversal of a baseline trend or the initiation of a prolonged (or at least semiprolonged) change in the intervention phase. When random variability is low, the first-order autoregressive model has no large random components to seed the sustained change process, and fewer confusions should occur. This was indeed what we found. Our explanation seems very appealing because it simultaneously satisfies several requirements. It is consistent with the mathematical nature of the first-order autoregressive model. It is able to account for the low error rate obtained when there was low random variation and positive autoregression. It is able to account for high false alarm rate given a conjunction of positive autocorrelation and higher random variation. Finally, it is able to predict the relatively lower false alarm rate obtained when there is random variation without the inertial effect of positive autocorrelation.

The discussion so far has focused on the high false alarms obtained. However, it is worth emphasizing that the detection of effects was generally good when effects were present. The implied corollary is that interjudge reliability for cases in which an effect was present was also good. Thus, even novice analysts seem capable of detecting realistically sized effects. It is, of course, possible that the "file-drawer" effect (Rosenthal, 1979) has distorted somewhat the effect size estimates; however, the effect sizes employed are clearly representative of at least a very substantial portion of behavioral cases.

A significant inverse relationship was found between the overall false alarm and miss rates among different judges. That is, individuals with higher false alarms tended to show lower miss rates and vice versa. This suggests that in a proper signal detection analysis, response bias is likely to be of actual rather than just potential importance. Response bias should be manipulable by incentive variations or instructions, and the effect size can be altered to encompass smaller signals. Thus, we envisage that investigations of the human operator characteristics (e.g., McNicol, 1972) are readily possible as well as desirable. The extent to which the high false alarm rate can be attributed to high operator noise or to response bias has practical implications for training programs and is not merely a question of theoretical interest. In the event that the problem is simply one of response bias, there seem to be reasonable prospects for improving performance by cognitive training. In the event of high noise in the human operator, the training problem may be much more complex.

The present study allowed five categories of response: no intervention effect, level change, trend change, combined level and trend change, and other type of systematic change. It is possible that, because four of the five response categories referred to an effect type, subjects may have experienced a bias towards an increased rate of effect responses. This may have contributed to the high false alarm rate. The present categories were employed because they represent those taught in the standard literature (e.g., Kazdin, 1982), and some texts include an even more elaborate set (e.g., Glass et al., 1975). Given this, it is not at all clear whether to interpret any tendency towards a "yes" response as a bias or merely a reflection of the way in which practitioners are asked to make decisions about single-case charts. Furthermore, it is possible that multiple effect categories are the ecologically valid choice and that our response categories represent those used implicitly by practitioners. However, this is an empirical question, and no work has as yet been directed to this issue. In conclusion, it seems unlikely that the extreme false alarm rates we observed were simply a consequence of having more response categories than yes/no, and it may be unreasonable to interpret such an effect as a bias even if it did occur.

The results of the present study have a number of important implications for research aimed at the development of a valid and viable single-case methodology for routine applied practice. Clearly, the replication of the study with other populations who have different levels of experience or educational background is urgent. Studies of the human operator, which investigate the noise and response bias characteristics, as well as variables that might be able to manipulate those characteristics are also suggested. Investigations of the effects of other case data on the decision process based solely on the chart, and investigation of visual aids that might improve the chart judgment, are required. We have just completed some work on the former, and others have commenced study of the latter (Hojem & Ottenbacher, 1988; Knapp, 1983). The effect of cognitive and perceptual training on the analysis of single-case time series also appears urgent in the light of the high false alarm rate reported here and the poor interjudge reliability found by other studies.

Routine use of statistical decision aids may have to be considered. Parsonson and Baer (1986), among others, have criticized the use of statistical methods as decision aids because they might encourage the acceptance of small effects and because they are not practical in the field. However, statistical methods do control the false alarm rate, unlike our sample of performers. If the high false alarm rate proves to be more generally typical of clinicians, the use of statistical decision aids, far from holding the danger of encouraging unnecessary liberalism, may be the way to prevent the apparently natural liberal bias of human judges. Of course the probability of a miss with our data is likely to be high for statistical models operating at $\alpha = 0.05$, given the brevity of the time series. Clearly, the conjoint comparison of false alarm and miss rates in human operators and statistical models subjected to the same data is another required investigation. We believe that the impracticability argument against statistical methods is overstated in the light of the desktop computer revolution and the potential to develop user-friendly software. Perhaps more serious questions are those that relate to the valid application of statistical methodology (Huitema, 1986b), particularly in the case of brief times series, such as occur in clinical practice. However, developments in the application of ITSA with approximate models that do not require prior model identification (Gottman, 1981; Velicer & McDonald, 1984) may be able to overcome previous difficulties. This is because the objection against the application of ITSA on the grounds that correct model identification requires a large sample (Huitema, 1986b), and the objection that model identification contains complexities unlikely to be mastered by practicing clinicians (Parsonson & Baer, 1986), are both bypassed by analysis without the step of model identification. The objection concerning small effect sizes may be overcome by defining effect size indexes and developing a normative model, as indeed others are already arguing in the clinical significance debate (Christensen & Mendoza, 1986; Jacobson, Follette, & Revenstorf, 1984, 1986; Wampold & Jenson, 1986). In any case, our findings suggest that statistical models will probably be more conservative than human judges.

The visual judgment literature to date, including the present study, has been confined to AB and ABAB designs. It is true that the AB comparison is a fundamental block in the more complex decision making of full designs, including the multiple baseline design. Indeed, the basic two-phase comparison is probably even more important than previously suggested in the literature, given the context of the a posteriori method of case management in which the appearance of the chart acts as a guide to revisions of the initial case design, particularly with respect to phase duration decisions (e.g., Kazdin, 1982). However, design issues require much more research than has hitherto appeared. The effects obtained with multiple baseline, changing criterion, and other clinically useful designs remain unknown. The possibility of accurate analysis of case data, be it achieved through improved visual aids, statistical aids, or special training programs for visual analysts, is not a goal readily abandoned, given the need for effective case methodology and the improvements already introduced by time-series designs.

REFERENCES

Barlow, D. H., Hayes, S. C., & Nelson, R. O. (1984). *The scientist practitioner: Research and accountability in clinical and education settings.* New York: Pergamon Press.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology,* **74,** 271–280.

Christensen, L., & Mendoza, J. L. (1986). A method of assessing change in a single subject: An alteration of the RC index. *Behavior Therapy,* **17,** 305–308.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press.

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

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

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

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

Gottman, J. M. (1981). *Time-series analysis: A comprehensive introduction for social scientists.* Cambridge, England: Cambridge University Press.

Hersen, M., & Barlow, D. H. (1976). *Single case experimental designs: Strategies for studying behavior change.* Oxford, England: Pergamon Press.

Hogarth, R. (1980). *Judgement and choice: The psychology of decision.* Chichester, England: John Wiley.

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

Huitema, B. E. (1985). Autocorrelation in applied behavior analysis: A myth. *Behavioral Assessment,* **7,** 107–118.

Huitema, B. E. (1986a). 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.

Huitema, B. E. (1986b). Statistical analysis and single-subject designs: Some misunderstandings. In A. Poling & R. W. Fuqua (Eds.), *Research methods in applied behavior analysis: Issues and advances* (pp. 209–232). New York: Plenum Press.

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

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1986). Toward a standard definition of clinically significant change. *Behavior Therapy,* **17,** 308–311.

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.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology,* **3,** 430–454.

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

Keppel, G. (1982). *Design and analysis: A researcher's handbook* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Knapp, T. J. (1983). Behavior analysts' visual appraisal of behavior change in graphic display. *Behavioral Assessment,* **5,** 155–164.

Kratochwill, T. R. (Ed.). (1978). *Single-subject research: Strategies for evaluating change.* New York: Academic Press.

Lichtenstein, S. C., Fischhoff, B., & Phillips, L. (1977). Calibration of probabilities: The state of the art. In H. Jungermann & G. de Zeeux (Eds.), *Decision making and change in human affairs* (pp. 275–324). Dordrecht, Holland: D. Reidel Publishing Company.

Marascuilo, L. A., & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences.* Monterey, CA: Brooks/Cole.

Matyas, T. A., & Greenwood, K. M. (1985). *A survey of serial dependence in behavioral baselines.* Paper presented at the 8th National Conference of the Australian Behaviour Modification Association, Melbourne.

Matyas, T. A., & Greenwood, K. M. (1990). *Problems in the estimation of autocorrelation in brief time-series and some implications for behavioral data.* Manuscript submitted for publication.

Matyas, T. A., & Greenwood, K. M. (in press). The effect of serial dependence on visual judgment of single-case charts: An addendum. *The Occupational Therapy Journal of Research.*

McNicol, D. (1972). *A primer of signal detection theory.* Sydney: Australasian Publishing Company.

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. (1978). The analysis and presentation of graphic data. In T. R. Kratochwill (Ed.), *Single-subject research: Strategies for evaluating change* (pp. 101–165). New York: Academic Press.

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 Press.

Rosenthal, R. (1979). The "file drawer" problem and tolerance for null results. *Psychological Bulletin,* **86,** 638–641.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1977). Behavioral decision theory. *Annual Review of Psychology,* **28,** 1–39.

Tversky, A., & Kahneman, D. (1971). The belief in the law of small numbers. *Psychological Bulletin,* **76,** 105–110.

Velicer, W. F., & McDonald, R. P. (1984). Time series analysis without model identification. *Multivariate Behavioral Research,* **19,** 33–47.

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

Wampold, B. E., & Jenson, W. R. (1986). Clinical significance revisited. *Behavior Therapy,* **17,** 302–305.

White, O. R. (1974). *The "split-middle" a "quickie" method of trend estimation.* Seattle: University of Washington, Experimental Education Unit, Child Development and Mental Retardation Center.

Williams, E. A., & Gottman, J. M. (1982). *A user's guide to the Gottman-Williams time-series analysis computer programs for social scientists.* Cambridge, England: Cambridge University Press.

Received January 10, 1989
Initial editorial decision June 30, 1989
Revisions received November 9, 1989; February 28, 1990
Final acceptance March 4, 1990
Action Editor, Terry J. Page