New Zealand Journal of Psychology, 1983, *12*, 69-73

The Visual Analysis of Time Series Data: Issues affecting the assessment of Behavioural Interventions[1]

Keith D. Ballard

University of Otago

The visual analysis of graphed data is valuable for assessing behavioural interventions because strong effects are necessary to produce obvious changes in the level or slope of the data display, reducing the probability of Type I errors. Visual inference may, however, be unreliable where autocorrelation is a property of the data. This aspect of behavioural data can be taken into account by appropriate statistical evaluations. Problems involved in both the visual and statistical analysis of behavioural data are examined in this paper.

Visual analysis of graphic data has been a basic procedure in applied behaviour analysis, allowing "close and continuing data contact" during the experiment (Parsonson & Baer, 1978) and providing others with access to all the primary data for independent interpretation and judgement. Parsonson and Baer (1978) have emphasised the value of the relative insensitivity of graphs compared with statistical analysis. With graphed data, differences between baseline and intervention data must be clearly evident for a convincing demonstration of an experimental effect. Statistical analysis, on the other hand, is more likely than graphic analysis to identify subtle effects as "significant", increasing the probability of a Type I error and thus affirming that a variable is a functional one when it is not (Baer, 1977).

Parsonson and Baer (1978) suggest that identifying only powerful variables is important, since only such variables are likely to have generalisable effects, thus encouraging the development of an effective and widely applicable technology. Less robust variables that must be identified by more indirect means are less likely to make a contribution to an applied analysis of behaviour.

Graphic analysis, however, is not without either problems or critics. The present paper reviews challenges to the efficacy and reliability of visual analysis of behavioural data, and considers evidence regarding the contribution statistical analysis could make to assessing behavioural interventions.

Problems in making visual inferences from graphed data

A visual evaluation compares performance during intervention with performance recorded in, and projected from, a baseline phase. Visual analysis can be made difficult by trends or excess variability in baselines, and by low magnitude of change or delays in behaviour change following intervention (Kratochwill, 1978). Interpretation is most readily defended when there are stable baselines, (or, if a trend is present it is in the opposite direction to that desired) and intervention performance either does not overlap with that occurring in the baseline phase or shows a trend in the desired direction that is not present, or is opposite to, the trend in baseline (Kazdin, 1978).

In addition to problems involved in the visual inspection of data identified above, some critics have argued against the sufficiency of "eyeballing" graphs and in support of a statistical analysis of data from time series experiments. Gottman and Glass (1978) for example, suggested that small effects rather than gross changes in behaviour might be expected in natural settings where there are complex patterns of stimuli, few of which can be controlled. Kazdin (1976) has

[1]Grateful appreciation is extended to Larry Nelson. and Terry Crooks for their comments on earlier drafts of this paper. Requests for reprints should be sent to Keith D. Ballard, Department of Education, University of Otago, P.O. Box 56, Dunedin, New Zealand.

suggested that failure to achieve dramatic changes need not mean an intervention is unimportant for applied purposes, and that variables exerting subtle control over behaviour should be investigated to see if their effects might be increased. Both Gottman and Glass (1978) and Kazdin (1976) suggest that while small effects might ultimately be rejected the researcher should at least identify them, if necessary using "sensitive" statistical procedures. Nevertheless, there is also the point that small effects might be a function of poor intervention strategies or poor research designs (Note 1). Researchers interested in less robust variables clearly must take account of such a possibility.

The most challenging argument against "eyeballing" graphs however, suggests that the visual analysis of time series data is not only unreliable but is inappropriate where autocorrelation is a property of the data. Observations that are repeated measures of human performance have a characteristic dependency in that successive observations in a time series tend to be correlated (McCain and McCleary, 1979). Knowing the level of performance of a subject at a given time allows predictions about subsequent points in the series (Kazdin, 1976). Correlation between data points separated by different time intervals is termed autocorrelation (Kazdin, 1976). Autocorrelation refers to a relationship of data points one to another such that in behavioural research at least, performance over time reflects reaction both to experimental conditions and to previous performance or learning (Note 2)

Two problems have been identified with autocorrelated data. First, except where baseline and intervention phase performance are quite distinct and non-overlapping, visual interpretation by experienced judges shows only modest inter-judge agreement (De Prospero & Cohen, 1979; Jones, Weinrott & Vaught, 1978). Second, when visual inferences have been compared with statistical analysis of less clear data, low levels of agreement have been shown (De Prospero & Cohen, 1979; Gottman & Glass, 1978; Jones, Weinrott & Vaught, 1978). From such research it has been argued that serial dependency within data does not allow reliable or valid inferences to be made from a visual analysis (Gottman & Glass, 1978; Note 2).

Statistical procedures for analysing behavioural time series data

An alternative, or supplement, to visual analysis is a statistical evaluation of intervention effects. However, simply projecting the baseline trend into the intervention phase and testing the significance of the difference of the predicted and obtained mean (e.g. Cameron & Robinson, 1980) is inappropriate. Serial dependency, while it does not bias estimates of the mean, biases estimates of the error variance (Kazdin, 1976; Hartman, Gottman, Jones, Gardner, Kazdin & Vaught, 1980) and in violating the assumption of independence of error components invalidates the use of t and F statistics (Kazdin, 1976; McCain & McCleary, 1979). This problem is resolved in interrupted time series analysis (ITSA). Here a model is identified for the error components of time series observations so that the systematic part of the error can be subtracted from each observation (Hartman, Gottman, Jones, Gardner, Kazdin & Vaught, 1980). The resulting scores, called residuals, contain no serial dependency and a change in mean across phases, reflected either as a change in level or slope (or both) can be tested for statistical significance (Kazdin, 1976).

Advocates of such a statistical approach argue that in contrast to the variability of opinions likely from a visual analysis, statistical analysis always gives the same result and is therefore more "reliable" (De Prospero & Cohen, 1979; Sharpley, 1981) The literature, however, suggests this may not be the case and identifies problems associated with the application of ITSA to behavioural data. ITSA requires, for example, the researcher to make judgements regarding the appropriate stochastic model (relating to the error components of observations) upon which subsequent analysis of intervention effects is based (Glass, Willson & Gottman, 1975; Hartman, Gottman, Jones, Gardner, Kazdin & Vaught, 1980; Sharpley & Rogers, Note 3). Clearly, therefore, the procedures are not free of opinion or judgement. Also, a major problem centres around the number of data points desirable or necessary before data can be sensibly evaluated using ITSA.

Sharpley and Rogers (Note 3) point out

that when the number of observations is small the standard error of the autocorrelation and partial autocorrelation coefficients (used in identifying the appropriate stochastic model) may be magnified and the estimates unreliable, so that model identification based on these estimates cannot be done with any confidence (see also Hartmann, Gottman, Jones, Gardner, Kazdin & Vaught, 1980). Glass, Willson and Gottman (1975) recommended at least 50 data points in each phase and others 50-100 observations in each phase (Box & Jenkins, 1970; Gottman & Glass, 1978). While the number of data points that might reasonably be analysed has been suggested as low as 20 (Hartmann, Gottman, Jones, Gardner, Kazdin & Vaught, 1980), advocates of ITSA present data from studies with far fewer observations, Gottman and Glass (1978), for example, analysing a study with seven base-line points and Jones, Vaught and Weinrott (1977, p. 161), reporting a "significant change in level" with only three data points in an intervention phase. This is far removed from the position that it is difficult to use ARIMA modelling (the autoregressive integrated moving average models used to assess the stochastic component of time series data—see Box and Jenkins, 1976) with fewer than 50 to 100 observations (McCain & McCleary, 1979).

Visual analysis vs statistical analysis of behavioural interventions

The problems posed by the preceding discussion involve first, the reliability of visual analysis of serially correlated data, especially where strong treatment effects are not evident. And second, the reliability and validity of interrupted time series analysis, especially with relatively small numbers of observations in each phase.

In regard to the latter problems, it is difficult to evaluate the contribution ITSA could make to behavioural studies until the issue of the number of data points necessary for reliable modelling is resolved. At present, the validity of modelling with less than 50 data points seems seriously in doubt. This alone limits the potential of ITSA for use in much applied behavioural research. The question then to be asked is how much of a problem does autocorrelation present for the visual analysis of time series data. In addressing this problem, a case will be made for the reliability and validity of visual analysis.

Visual analysis usually involves inspection of the primary observation data from a study. Identification of treatment effects is on the basis of clear changes in the level and/or slope of performance trends following intervention. A valid inference that behaviour change resulted from intervention can be made when the level or slope of data in the intervention phase is obviously and substantially divergent from that pertaining during baseline. Also, it is important to note that such inferences are not properly made on the basis of one comparison between a baseline and a treatment phase, but should involve a replication of the treatment effect in a reversal, withdrawal, multiple baseline or related within subject design (Hersen & Barlow, 1976; Wampold & Furlong, 1981).

As mentioned above, where performance changes are not distinct and non-overlapping the inter-judge reliability of visual analysis is in question. Clearly this presents a significant problem in evaluating research where, at best, a weak effect may be apparent. Nevertheless, such a problem, while not to be ignored, does not present an overwhelming challenge to the efficacy of visual analysis of graphic data. Assuming that both the measurement processes and research design of a study were sound, then the sensible response to marginal experimental effects is to question the efficacy of the intervention. A decision would then be made regarding either a replication of the study to provide further evaluation, or to reject the intervention as of no applied value.

Applied behaviour analysis has typically been concerned with powerful variables that will have socially meaningful effects in natural settings. Such concerns tend to protect against assigning efficacy to variables that have only a marginal impact on behaviour (Baer, 1977). The knowledge that evaluating marginal changes across experimental phases may be unreliable is a further protection against Type I errors in visual analysis.

Peer review of research submitted to behavioural journals is one example of visual analysis of graphic data in natural settings. Even if editors and reviewers are

collectively in error, journal readers should have the primary research data displayed for their own appraisal. It would be unfortunate if applications of ITSA included a trend away from graphic presentations and reliance on summary data and tests of statistical significance.

Finally, there is a pressing need for research into the processes used by behaviour analysts in the visual evaluation of graphed data. Wampold and Furlong (1981) have shown that persons trained in visual inference attend primarily to large differences in the data in different phases, while persons with statistical training take more account of relative variation in the data. Such research has imporant implications for training in behaviour analysis.

Conclusions

Applied behavioural research has primarily been concerned with identifying variables associated with obvious and meaningful change in human behaviour. Allegiance to visually clear changes in graphic data displays has been seen as resulting in few Type I errors. Caution regarding apparently marginal experimental effects has probably resulted in many Type II errors. Critics of the visual analysis of interrupted time series data have been concerned mainly with instances where data in adjacent experimental phases is overlapping and without distinct changes in level or slope. In such cases the autocorrelational properties of the data make inferences based on visual analysis unreliable.

While ITSA may be a useful form of additional analysis where there are apparently marginal experimental effects, these statistical procedures are not free of subjective judgement, and at present are recommended for cases where there are at least 20 to 50 data points in each phase, numbers not frequently encountered in applied behavioural research.

Research is needed into the processes of visual inference and into training in the visual analysis of behavioural data.

Reference Notes

1. Glynn, T. Personal communication, 27 August, 1981.
2. Sharpley, C. *Visual analysis of operant data: Can we believe our eyes?* Paper presented at the Third Conference of the Australian Behaviour Modification Association, Melbourne, May, 1981.
3. Sharpley, C. & Rogers, H. J. *Means, graphs, ts and Fs: Inadequate measures in operant research?* Unpublished manuscript, Centre for Behavioural Studies, University of New England, Armidale, N.S.W., Australia, 1981.

References

Baer, D. M. Perhaps it would be better not to know everything. *Journal of Applied Behaviour Analysis,* 1977, *10,* 167-172.

Box, G. E., & Jenkins, G. M. *Time-series analysis: Forecasting and control.* (2nd ed.). San Francisco: Holden-Day, 1976.

Cameron, M. I. & Robinson, V. M. Effects of cognitive training on academic and on-task behaviour of hyperactive children. *Journal of Abnormal Child Psychology,* 1980, *8,* 405-419.

De Prospero, A., & Cohen, S. Inconsistent visual analyses of itrasubject data. *Journal of Applied Behavior Analysis,* 1979, *12,* 573-579.

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

Gottman, J. M., & Glass, G. V. Analysis of interrupted time-series experiments. In T. R. Kratochwill (Ed.), *Single subject research: Strategies for evaluating change.* New York: Academic Press, 1978.

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

Hersen, M., & Barlow, D. H. *Single case experimental designs: Strategies for studying behavior change.* New York: Pergamon Press, 1976.

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

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

Kazdin, A. E. Statistical analyses for single-case experimental designs. In M. Hersen & D. H. Barlow (Eds.), *Single case experimental designs: Strategies for studying behavior change.* New York: Pergamon Press, 1976, pp. 265-316.

Kazdin, A. E. Methodology of applied behavior analysis. In A. C. Catania & T. A. Brigham (Eds.), *Handbook of applied behavior analysis: Social and instructional processes.* New York: Irvington Publishers, 1978, pp. 61-103.

Kratochwill, T. R. Foundations of time-series research. In T. R. Kratochwill (Ed.), *Single subject research: Strategies for evaluating change.* New York: Academic Press, 1978, pp. 1-100.

McCain, L. J., & McCleary, R. The statistical

analysis of the simple interrupted time-series quasi-experiment. In T. D. Cook & D. T. Campbell (Eds.), *Quasi-experimentation: Design and analysis issues for field settings. Chicago:* Rand McNally College Publishing Co., 1979, pp. 233-294.

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

Sharpley, C. Time-series analysis of counselling research. *Measurement and evaluation in guidance,* 1981, *3,* 149-157.

Wampold, B. E., & Furlong, M. J. The heuristics of visual inference. *Behavioural Assessment,* 1981, *3,* 79-92.

The following people acted as reviewers of manuscripts submitted for publication in the New Zealand Journal of Psychology volume 12, Numbers 1 and 2. The Editors and members of the Editorial Board gratefully acknowledge their assistance.

B. Barrer
N. Blampied
P. Bull
J. Bushnell
J. Clarkson
S. Cullen
G. Fletcher
A. Forbes
J. Gribben
B. Heskyth
D. Hughes
B. Jamieson
R. Kammann
B. Keeling
S. Kemp
N. Long
I. McCormick
K. McFarland
S. Ng
M. O'Driscoll
D. O'Hare
B. Parsonson
J. Ritchie
B. Stacey
B. Tuck