JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1978, 11, 277-283 NUMBER 2 (SUMMER 1978)

EFFECTS OF SERIAL DEPENDENCY ON THE AGREEMENT BETWEEN VISUAL AND STATISTICAL INFERENCE[1]

RICHARD R. JONES, MARK R. WEINROTT, AND RUSSELL S. VAUGHT

EVALUATION RESEARCH GROUP AND SUNY, BINGHAMTON

Comparisons between visual and time-series inferences from behavorial data show that serial dependency in scores is likely to disrupt agreement between the two methods of analysis. If researchers follow an earlier recommendation that time-series analysis be used to supplement or confirm visual analysis, this study's findings suggest that the two methods will disagree most often when the data contain high levels of autocorrelation and when reliable behavorial changes are indicated by time-series analysis.

DESCRIPTORS: statistics, time-series analysis, inferences-visual *versus* statistical

When analyzing individual-subject experiments, behavior analysts make inferences about behavioral change based, usually, on visual analysis of the data. For example, comparisons between a baseline phase and an intervention phase are made on the basis of visually apparent increases or decreases in scores for a target behavior. Inferences based on statistical comparisons of changes in level have been eschewed by many operant researchers (Baer, 1977; Michael, 1974), although recent proponents of statistical methods for analysis of single-subject experiments have been heard (*e.g.,* Gentile, Roden, and Klein, 1972) and criticized (Hartmann, 1974).

Whether visual or statistical methods are used as judgemental aides (Michael, 1974) in analyzing operant experiments, a basic question involves the agreement between inferences made from different analyses of the same data. This paper reports a study that compared visual inferences about changes in operant experiments with inferences based on time-series analysis of the same experiments. Additionally, the study was designed to appraise the influence of serial dependency in behavioral scores on the agreement between inferences based on visual analysis or time-series analysis.

Time-series analysis and the concept of serial dependency were reviewed nontechnically in Jones, Vaught, and Weinrott (1977). To conserve space, only a brief summary of time-series analysis and serial dependency is provided here. However, to understand fully the study reported below, the reader should carefully review the earlier article.

Time-series analysis is a statistical procedure for making inferences about changes in level and trend among the several phases of an individual-subject experiment. Time-series analysis accommodates a common property of temporally ordered behavioral scores, serial dependency, which violates the independence assumption underlying more common statistical methods (*e.g.,* analysis of variance). Serial dependency means simply that temporally adjacent scores tend to be related to, or predictive of, one another. For example, an individual subject's score for Day 1 tends to predict the subject's score on Day 2. A Day-2 score will tend to predict the Day-3 score, *etc.* Jones *et al.* (1977) discussed serial dependency and noted

[1]This study was supported by Grant No. R01 MH 25631 and R01 MH 25548 from the Center for Studies of Crime and Delinquency, National Institute of Mental Health, U.S. Public Health Service. Reprints may be obtained from Richard R. Jones, Evaluation Research Group, 220 E. 11th Avenue, Eugene, Oregon 97401.

that it is a prevalent property of individual subject scores, and must be accommodated by any statistical procedure applied to such data.

Since serial dependency is a problem for making inferences from statistical analysis of temporally ordered scores, it may also be a problem for making inferences based on visual analysis of individual subject scores in operant experiments. The present study was designed to determine whether or not serial dependency influences the agreement between inferences based on visual or time-series analysis.

METHOD

Subjects

To study the agreement between visual and time-series inferences, and to test for the effects of serial dependency on agreement, *JABA* graphs were presented to a panel of 11 judges who were familiar with operant experiments. Judges were full-time researchers, university professors, and graduate students with 3 to 17 yr of research experience in psychology, including applied behavior analysis. Each judge was asked to decide whether or not a meaningful change in level was demonstrated from one phase to another in each of the graphs. "Meaningful" referred only to the reliability of the change and not to its social value.

Stimulus Materials

Published experiments were sampled from *JABA* using the following selection criteria. First, experimental effects claimed by the authors and depicted in the graphs had to be sufficiently nonobvious to warrant critical analysis. Second, studies were chosen that used multiple baselines, several different phases, small numbers of data points within phases, and unequal numbers of data points across phases. Third, attention was given to experiments where serial dependency might be evidenced by possible nonzero trend, apparent from visual inspection of the graphs. The following partial list of studies illustrates the variety of experiments chosen: (a) a single-component study—AB (Boren and Colman, 1970); (b) a traditional reversal design—ABAB (Ingham and Andrews, 1973); (c) a multiple component study—ABCB (Phillips, Phillips, Fixsen, and Wolf, 1971); (d) a multiple-baseline study—A/B/C/B/C (Baer, Rowbury, and Baer, 1973); and (e) a reversal component study—ABACADEA (Wincze, Leitenberg, and Agras, 1972). Some of these and the other studies sampled involved combined data for a number of subjects; others involved individual subjects.

This selection of *JABA* studies obviously produced a nonrandom sample of operant experiments. Random sampling of studies was considered, but rejected for the following reasons. First, a random sample would have contained many studies showing dramatic behavioral changes, where no reasonable critic would disagree with the inference(s) drawn by the author(s). For such studies, agreement between inferences based on this study's judges' visual analysis and the time-series analysis would have been high. But dramatic effects probably are properly interpreted using visual inference, and time-series analysis as a supplementary method for drawing inferences would not be recommended in such studies (Jones *et al.,* 1977). So, for large-effect studies, there seems to be no issue about agreement between visual and time-series inferences, since the latter probably are unnecessary for adequate interpretation of the data. Hence, the random selection of such experiments for this study would have been of little interest, since time-series analysis would not typically be used for such experiments. This was the reason for the first selection criterion discussed above, which excluded large-effect experiments from this sample.

The final sample of studies included 24 different experiments, each depicted in a graph as originally published in *JABA.* Twenty of the 24 experiments (83%) had statistically significant lag 1 autocorrelations ($p < 0.05$). The range of these significant lag 1 autocorrelations was from 0.40 to 0.93. Nine of the 20 significant autocorrelations were greater than 0.70. Thus, serial dependency, as measured by autocorrelations, appears to be a relatively common property of behavioral scores used in this sample of operant experiments.

A total of 85 test items were obtained from the 24 graphs, but only 58 of these were used in this study, as described below. A test item was a pair of adjacent phases, for each of which subjects judged the meaningfulness of change in level. For example, an $A_1B_1A_2B_2$ experiment yielded three test items, A_1B_1, B_1A_2, and A_2B_2. Each graph, containing from one (AB) to seven test items (ABACADEA), was projected on a screen in clear view of the 11 judges. The graph contained all of the published information provided by the author(s), as the graphs were simply reproduced directly from the pages of *JABA.*

Agreement Analysis

The judges rated the test items for change in level, and their responses were coded as "yes" (meaningful change in level), "unsure" (indeterminant), or "no" (not meaningful change in level). The same test items were analyzed by formal time-series methods and the change in behavior was coded as statistically significant ($p < 0.05$) or statistically nonsignificant ($p > 0.10$). Agreements and disagreements between the judges' visual inferences and the time-series inferences were obtained as follows. Agreement was scored if visual inference was coded "yes" and time-series analysis was significant ($p < 0.05$) or if visual inference was coded "no" or "unsure" and time-series analysis was nonsignificant ($p > 0.10$). Disagreement was scored if visual inference was coded "yes" and time-series analysis was nonsignificant or if visual inference was coded "no" or "unsure" and time-series analysis was significant ($p < 0.05$). The usual formula for agreement proportions was used [*i.e.*, $P_A = A/(A + D)$] to obtain an agreement index for each judge over the test items.[2] These agreement indices were then used as dependent variables in the analysis of variance design presented below. Averaging over all 58 testing items, the agreement indices for the 11 judges ranged from 0.50 to 0.65. Note that 0.50 represents chance agreement. Hence, the best agreement between visual inferences and time-series inferences was only 15 percentage points above chance.

Design

The 24 graphs were partitioned into three sets of eight each, on the basis of their serial dependency. Lag 1 autocorrelation coefficients were computed over all data points in all phases for each of the 24 graphs. The graphs were then ranked on the lag 1 autocorrelation and assigned to one of three sets. The "low" set was composed of eight graphs with autocorrelations ranging from 0.15 to 0.50 (14 test items), the "moderate" set was composed of the eight graphs with autocorrelations from 0.51 to 0.75 (25 test items), and the "high" set was composed of the remaining eight graphs with autocorrelations from 0.76 to 0.94 (19 test items). These three categories of test items comprised the serial dependency factor for analysis of the agreement indices.

Factor B was labelled "significance level"; that is, the test items were classified as to the presence or absence of a significant change in level as determined statistically by time-series analysis (Jones *et al.*, 1977). The test items were classified into three sets: (a) 35 items where the *t*-test was statistically significant ($p < 0.05$); (b) 35 items where the *t*-test was not statistically significant ($p > 0.10$); and (c) 13 items where the *t*-test significance level fell between 0.05 and 0.1. These latter 13 items were discarded from the analysis to ensure only the inclusion of items where change in level was unambiguously significant or nonsignificant.

These classifications of the test items by serial dependency (Factor A) and significance of

[2] The "unsure" responses accounted for only 6.8% of all judgements, so it was not considered problematic for this agreement analysis to combine "unsure" responses with "no" responses.

change in level (Factor B) produced an inordinately large set of items (28) with moderate serial dependency and nonsignificant level changes. To equate more nearly the numbers of test items from which accuracy scores were obtained in each cell of the design, half of these 28 items were randomly dropped from the analysis. The remaining 14, and all test items classified into each of the other five cells of the design, were included in the study.

The dependent variable for the two-factor, repeated measures ANOVA was the agreement index for each judge for each of the six sets of items. For example, if the judge and the time-series analysis agreed on seven of the nine test items in the upper-left cell of Table 1, an agreement index of 7/9 or 0.78 would be obtained. The 66 agreement indices, 11 for each of the six cells, were treated as within-cell replications in the design. Since the same subjects were used to judge all test items, the agreement indices in one cell of the design could not be considered independent of those in another cell. Thus, a repeated measures ANOVA design was used, with repeats on both the serial dependency and significance level factors.

RESULTS

A significant main effect for Factor A (serial dependency) showed that differences in mean agreement varied as a function of serial dependency in the test items ($F = 20.88$; $df = 2, 20$; $p < 0.001$). As expected, agreement between visual and time-series inferences was inversely related to the magnitude of the serial dependency in the scores. For test items where serial dependency was low, mean agreement was 0.73 between the judges and time-series analysis, as opposed to 0.54 for the medium level and 0.50 for the highest level of serial dependency. Newman-Keuls comparisons (Winer, 1971) among the three pairs of means showed a significant difference ($p < 0.02$) between the mean agreement for the low autocorrelated items (0.73 agreement) and both the moderate (0.54) and high autocorrelated items (0.50), but no difference between the latter two means.

Table 1

Mean agreements between visual inferences and time-series inferences, over judges for test items classified on serial dependency and significance factors.

Factor B Significance level	Factor A Serial Dependency			
	Low	Moderate	High	
$p < 0.05$	0.58	0.51	0.48	0.52
$p > 0.10$	0.89	0.60	0.52	0.67
	0.73	0.55	0.50	0.60

A nearly significant main effect ($F = 4.61$; $df = 1, 10$; $p < 0.055$) for Factor B (significance level) was particularly interesting because the direction of the difference between the mean agreements was not expected. Agreement was greatest (0.67) for test items that were *not* statistically reliable by time-series analysis, and poorest (0.52) for items that *were* statistically reliable. That is, visual and time-series inferences agreed better when the statistical test indicated nonsignificant changes in level than when significant changes in level were indicated. This suggests that statistically reliable experimental effects may be more often overlooked by visual appraisals of data than nonmeaningful effects. If time-series analysis were used to supplement visual analysis, as proposed by Jones *et al.* (1977), researchers probably would infer meaningful changes in their data more often than if visual inferences alone were used to analyze operant experiments.

The interaction between serial dependency (Factor A) and significance level (Factor B) was also significant ($F = 6.90$; $df = 2, 20$; $p < 0.01$). Newman-Keuls comparisons showed that this interaction effect was attributable to high agreement on items that had the lowest autocorrelations and which showed no experimental effect from phase to phase. Visual and time-series inferences averaged 0.89 agreement on these items, a significantly ($p < 0.01$) higher mean agreement than for any other combination of the two factors. No differences were found between any other pairs of within-cell means for the interaction effect.

This interaction effect suggests that visual and time-series inferences agree best when the data show neither serial dependency nor meaningful experimental effects. Interestingly, these two conditions are either unlikely or unwanted in most operant research. Yet, it is under these conditions that agreement or reliability of inferences under the two methods of analysis is greatest. In contrast, agreement between visual and time-series inferences was lowest under the opposite set of conditions, namely, high serial dependency and statistically reliable changes in level ($x = 0.48$). And this combination of conditions is both likely and wanted in operant research. But, the agreement between visual and statistical inferences was less than chance under these conditions!

DISCUSSION AND CONCLUSIONS

It would seem that three general conclusions can be drawn from this study. First, visual inferences and time-series inferences did not agree particularly well—a mean agreement of 0.60 over all 58 test items is not much greater than chance agreement (0.50). Second, agreement between visual and time-series inferences varied reliably across the three levels of serial dependency. When serial dependency was high, agreement suffered, and when low, agreement was better. And third, when time-series tests suggested statistically reliable changes in behavior, agreement was relatively low. In contrast, when no reliable change in behavior was indicated by time-series analysis, visual and statistical inferences agreed better.

Only the second of these conclusions was expected. Serial dependency was hypothesized as a potential disruptor of agreement between visual and time-series inferences, and in fact this appears to be the case. So, when serial dependency is high in any given operant experiment, we can expect less agreement between inferences based on visual analysis and time-series analysis than when serial dependency is low or nonexistent. Unfortunately, the former case is more likely in operant research, because serial dependency appears to be a prevalent characteristic of experimental data in operant studies. So, if the recommendation of Jones *et al.* (1977) is followed, *viz.* that time-series analysis be used to supplement visual analysis, researchers can expect frequent disagreement between inferences that would be drawn from visual analysis and time-series analysis of the same data.

The first conclusion, that overall agreement between visual and time-series inferences was not much greater than chance agreement, was not expected. When two methods of data analysis disagree in the inferences to be drawn from them, one naturally wonders which is best. Is visual inference not to be trusted? Is time-series analysis not to be trusted? Or, is neither to be trusted? When agreement is low, as in this study, each method could be suspect.

However, if one does wish to choose between the two methods (rather than use both, as the writers have recommended), then at least two other sets of information deserve to be considered. First, each method's track record as a judgemental aid in behavioral research could be considered; and second, the reliability of each method could be considered. The first consideration is highly subjective and discussion of it tends quickly to polarize the issues unproductively. We prefer to avoid these polemics by discussing only the second consideration. So, what about the reliability of the inferences derived from the judges' visual analyses and the reliability of the inferences derived from the time-series analyses?

Consider the reliability of inferences derived from time-series analysis. These inferences are perfectly reliable, in that no matter who applies the time-series method, or how many different times to the same data, the same inference(s) will be drawn. Statistical inference is perfectly reliable in this mundane sense, but this perhaps is one of its advantages over visual inference.

Now, note that interjudge reliability was far from perfect. In fact, the data suggest a general failure of judges to arrive at a consensus. The intercorrelations (over the 58 test items) among the 11 judges' agreement scores ranged from 0.04 to 0.79, with a median of 0.39. An average interjudge reliability coefficient of 0.39 does not suggest a high consensus among judges, and casts doubt on the dependability of these visual inferences in this study.

But some judges were reliable and other were not, and this variability in judges' reliabilities raises the possibility that their agreement indices would be effected by their reliabilities. After all, if a judge performed the judgement task unreliably, how could that judge possibly show good agreement with the time-series results? To test this possibility, the rank order correlation (over the 11 judges) between the judges' (a) average reliabilities (phi coefficients) with the other judges and (b) agreement indices was calculated. The obtained rho −0.19 indicates virtually no relationship between interjudge reliability and agreement between the judges' visual inferences and time-series inferences. Hence, the level of a judge's agreement with the time-series inferences was not associated with the judge's reliability. Unreliable judges were just as likely to agree with the time-series inferences as were reliable judges.

So, we have a situation where visual inferences suffer from relatively low reliability (poor interjudge consensus), while time-series inferences do not, and the two sets of inferences do not agree very well. And further, the low agreement is not associated with poor judge reliability. Given this information, one probably should question the worth of the visual inferences.

Other criticisms of this study could be offered to explain away the finding of low agreement between the visual inferences and the time-series inferences, including (a) the training and qualifications of the judges and (b) the suitability of the inferential task required of the judges. Another sample of judges, say the *JABA* Board of Editors, might show higher interjudge reliability than obtained here. And, another sample of judges' inferences might show stronger agreement with the time-series inferences than obtained here. Also, a differently defined inferential task required of the judges might improve their reliabilities and their agreement with time-series inferences. While these alternatives might change the magnitude of agreements between the two methods of making inferences, and certainly should be tried in replications of this study, we will argue that the presence of serial dependency will still lower agreements between the two methods, even if reliabilities and overall agreements are raised.

Regardless of possible inadequacies in the present study's particular sample of judges or the inferential task required of them, note that some judges showed strong reliabilities with other judges and some judges showed strong agreement with the time-series inferences. In fact, the mean agreement of 0.73 between visual and time-series inferences for the low serially dependent items is rather respectable. And the mean agreement of 0.89 for the items in the low dependency and $p > 0.10$ significance cell is most satisfactory. So, visual inferences by judges can be reliable, and they can agree quite well with inferences derived from time-series analyses. What seems to be needed now is further study of other conditions under which agreement between visual inference and statistical inference varies. From this study, it seems clear that agreement between visual inference and time-series inference will be lowered when serial dependency exists in the data and when statistically reliable changes in behavior are indicated by time-series analysis. Since these conditions in behavioral data are, respectively, likely and desirable, probably time-series inferences would prove to be a useful supplement to visual inferences.

REFERENCES

Baer, A. M., Rowbury, T., and Baer, D. M. The development of instructional control over classroom activities of deviant preschool children. *Journal of Applied Behavior Analysis*, 1973, **6**, 289-298.

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

Boren, J. J. and Colman, A. D. Some experiments on reinforcement principles within a psychiatric ward for delinquent soldiers. *Journal of Applied Behavior Analysis*, 1970, **3**, 29-37.

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

Hartmann, D. P. Forcing square pegs into round holes: some comments on an analysis of variance model for the intrasubject replication design. *Journal of Applied Behavior Analysis*, 1974, **7**, 635-638.

Ingham, R. J. and Andrews, G. An analysis of a token economy in stuttering therapy. *Journal of Applied Behavior Analysis*, 1973, **6**, 219-229.

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

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

Phillips, E. L., Phillips, E. A., Fixsen, D. L., and Wolf, M. M. Achievement Place: modification of the behaviors of predelinquent boys within a token economy. *Journal of Applied Behavior Analysis*, 1971, **4**, 45-49.

Wincze, J. P., Leitenberg, H., and Agras, W. S. The effects of token reinforcement and feedback on the delusional verbal behavior of chronic paranoid schizophrenics, *Journal of Applied Behavior Analysis*, 1972, **5**, 247-262.

Winer, B. J. *Statistical principles in experimental design.* New York: McGraw-Hill Book Company, 1971.

Received 11 November 1976.
(Final Acceptance 3 November 1977.)