JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1977, 10, 151-166 NUMBER 1 (SPRING) 1977

TIME-SERIES ANALYSIS IN OPERANT RESEARCH[1]

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

OREGON RESEARCH INSTITUTE AND CENTER FOR CREATIVE LEADERSHIP

A time-series method is presented, nontechnically, for analysis of data generated in individual-subject operant studies, and is recommended as a supplement to visual analysis of behavior change in reversal or multiple-baseline experiments. The method can be used to identify three kinds of statistically significant behavior change: (*a*) changes in score levels from one experimental phase to another, (*b*) reliable upward or downward trends in scores, and (*c*) changes in trends between phases. The detection of, and reliance on, serial dependency (autocorrelation among temporally adjacent scores) in individual-subject behavioral scores is emphasized. Examples of published data from the operant literature are used to illustrate the time-series method.

DESCRIPTORS: ANOVA, experimental design, methodology, multiple baseline, serial dependency, single-organism research, statistics, time-series analysis

Reversal and multiple-baseline designs are the methodological kingpins in the functional analysis of behavior. These designs use a baseline period to assess the typical performance of a subject's target behavior, followed by an experimental manipulation intended to alter the level of the target behavior. In reversal designs, the intervention phase is followed by another baseline period to demonstrate the efficacy of the experimenter's control over the target behavior. In multiple-baseline designs, the intervention is implemented at different times to gauge its impact on each separate behavior or subject.

Behavioral scores are plotted on a time line running through the baseline and intervention phases, and the return-to-baseline phase in a reversal study. The temporal order of the behavioral scores is an intrinsic and unalterable characteristic of such time series, as is the temporal arrangement of the baseline, intervention, and return-to-baseline phases (Risley and Wolf, 1972).

The behavioral scores used in these two designs constitute an interrupted time series (Campbell and Stanley, 1970). The subjects' scores are displayed over time, with interruptions in the time series designated as the change points from one to another phase of the design. In the typical reversal design, interruptions occur at the transition between baseline and intervention phases, and again between the intervention and return-to-baseline phases. The main problem for analysis of such interrupted time-series data is to determine whether or not changes in the behavioral scores following the interruptions warrant the conclusion that experimental control over behavior has been obtained.

Different methods can be used to assess the effects of interruptions on behavioral scores in reversal or multiple-baseline studies. Operant researchers have typically relied on visual inspection of their data when drawing conclusions about the efficacy of their experimental interventions. Recently, although controversially, conventional analysis-of-variance models have been suggested for statistically comparing mean scores from each of the several phases in reversal designs (Gentile, Roden, and Klein, 1972; Hartmann, 1974; Thoresen and Elashoff, 1974).

[1]Supported by R01 MH 15985 and R01 MH 25631-01 from the National Institute of Mental Health, U.S. Public Health Service. Reprints may be obtained from R. R. Jones, P.O. Box 3196, Eugene, Oregon 97403.

The purpose of this paper is to recommend the use of a third procedure, time-series analysis, as a supplement to visual analysis in operant studies (*e.g.,* Schnelle and Lee, 1974). Time-series analysis has been developed for use in other disciplines (*e.g.,* econometrics and meteorology), and seems to have potential for the analysis of the effects of intervention in individual-subject, interrupted time-series designs. In what follows, arguments favoring time-series analysis are outlined, and the method is applied illustratively to data published in the operant literature. Before discussing the time-series method, the general matter of drawing inferences from data, whether in operant or other studies, should be addressed.

On Making Inferences from Nonindependent Scores

Analysis of reversal or multiple-baseline experiments in applied operant research is aimed at drawing conclusions about behavior change due to an intervention. Conclusions are based on human judgments or inferences about the impact of an intervention on some target behavior. Although they require human judgment, these inferences about behavior change are not based on subjective, personal criteria. Rather, inferences are made using criteria that are known to, or can be communicated to, other operant researchers. Examples of some criteria used in making inferences about behavioral change include the stability of baseline behavioral scores, the variability of behavioral scores within and across phases of an experiment, and the amount of overlap between scores from adjacent phases, *e.g.,* baseline and intervention phases of an experiment. Coupled with these criteria is the requirement that there are enough scores in each phase of an experiment to justify the inferences or conclusions drawn by the researcher. For example, most operant researchers would not be convinced by a study with only one score in each of three or more phases of the experiment. The number of scores per phase required to support inferences about behavioral change depends largely on the other criteria, *viz.,* the stability, variability, and between-phase overlap of the observed scores.

These criteria, which form the basis for making inferences about change in operant studies, are statistical in nature—*i.e.,* stability, variability, overlap, and numbers of scores are statistical concepts. Hence, it is clear that inferences or conclusions drawn by operant researchers from reversal or multiple-baseline studies require consideration of statistical properties of the experimental data. It is irrelevant here whether the operant researcher actually calculates statistical indices (*e.g.,* score variances) or visually appraises the statistical properties of behavioral scores. The point is that statistical properties of behavioral data form the basis for the operant researcher's inferences and conclusions about the impact of interventions in reversal or multiple-baseline experiments. Although it is not in the tradition of operant methodology to calculate statistical indices, or to conduct statistical inference tests, applied behavior analysts are skilled in visual appraisal of the statistical properties of behavioral data and in drawing inferences or conclusions about behavioral change from such appraisals.

In addition to the stability, variability, overlap, and number-of-scores properties of behavioral data in operant experiments, there is another statistical property of behavioral scores that heretofore has not been recognized as an important influence on inferences about behavioral change in reversal and multiple-baseline experiments. This property is serial dependency. It has been identified as the major reason why certain recently suggested statistical analyses of individual subject data are not appropriate (Hartmann, 1974; Thoresen and Elashoff, 1974). Operant researchers who have tended to eschew inferential statistical procedures may have been relieved to learn that serial dependency precludes the uncritical use of certain conventional statistical methods (*e.g.,* analyses of variance to compare means from two or more phases of an operant experiment). Unfortunately, serial dependency

not only biases the results of conventional inferential statistical procedures, but it also interferes with inferences about changes in behavior derived from the customary visual appraisal of statistical properties of operant data.

What is serial dependency and why does it interfere with visual inferences about behavioral change? Serial dependency is a common property of behavioral scores, such as repeated observations of a single subject. For example, if in a classroom study the proportion of time spent on-task is obtained each day for two weeks for a pupil, it is likely that the 10 temporally ordered scores in this series will not be independent of one another. That is, the score for Day 1 will be related to the score for Day 2, which in turn will be related to the score for Day 3, *etc.* When this occurs, the sequence of scores for the subject is said to be serially dependent. The term "serial" refers to the fact that the temporal order of the scores is an inherent and inviolable property of the scores, and the term "dependent" refers to the relationship between scores in the temporally ordered series.

Serial dependency in temporally ordered behavioral scores for a single subject (or for the mean score from a sample of subjects) is a statistical property of reversal or multiple-baseline data that is not as well known or understood as are other statistical properties, such as stability, variability, overlap among scores, or numbers of scores. Researchers are able to appraise these latter and better-known statistical properties by, for example, visually estimating trends (stability), determining the range of scores (variability), counting the number of scores in one phase of an experiment that fall within the range of scores from an adjacent phase (overlap), and counting the number of scores within a phase (number). In contrast, serial dependency in behavioral scores cannot be so handily appraised as these other statistical properties. But the importance of serial dependency in behavioral scores demands that operant researchers become familiar with it, and with a fairly simple procedure for appraising the extent of serial dependency in any given series of behavioral scores obtained in a reversal or multiple-baseline experiment.

Serial dependency is appraised by calculating a statistic called an autocorrelation coefficient. Existing computer programs[2] routinely calculate autocorrelations as one stage in the time-series analysis of behavioral scores, so the computational details will be excluded from this discussion. Instead, a conceptual description of autocorrelation is provided, and readers who are familiar with conventional correlation methods will readily see the similarity with autocorrelations.

An autocorrelation indicates the extent to which scores at one time point in a series are predictive of scores at another time point in the series. To obtain an autocorrelation, pairs of scores from the series are formed as follows. The score from time point 1 is paired with the score from time point 2, the score from time point 2 is paired with the score from time point 3, *etc.* When the pairs are formed with scores from adjacent time points, the resultant coefficient is called a lag 1 autocorrelation, since there is one time-point lag or difference between the two scores in each pair. Larger lags can be formed by pairing, for example, score 1 with score 3, score 2 with score 4, *etc.*, to obtain a lag 2 autocorrelation. The autocorrelation coefficient is interpreted exactly like the conventional correlation, except that the degrees of freedom for determining the significance of the coefficient are reduced by the number of lags. For example, in conventional correlation analysis, the *df* are $N - 2$, while for autocorrelation, they are $N - 2$ minus the number of lags (where N is the number of pairs of scores over which the correlation is obtained, and *lags* is the number of time points between the scores in each pair).

If the lag 1 autocorrelation for a series of

[2]Existing computer programs may be obtained by contacting their authors, *e.g.*, the Time Series Programs, TMS and CORREL, Bower, Padia, and Glass are available from Laboratory of Educational Research, University of Colorado, or from the first author.

behavioral scores from an operant experiment is statistically significant (say at the usual 0.05 level of confidence), then it can be said that the scores are serially dependent. Often, when the lag 1 autocorrelation is significant, the coefficients for larger lags will be also, but usually the size of the correlations decreases as the lags increase. This means that the within-subject predictiveness of the scores lessens as the time between scores lengthens, which makes good intuitive sense.

Serial dependency, as measured by autocorrelations, is quite common in behavioral scores for individual subjects. In fact, it could be argued that serial dependency should always be found in repeated measurements for individual subjects, unless, of course, the time intervals between scores are so large or so irregular as to preclude any reasonable expectations of predictability from one time point to later ones. The reason that one should expect serial dependency is simply that people and their environments do not behave or function randomly over time. If they did, then one could argue that the entire psychological enterprise (*i.e.,* the prediction and control of behavior) has been tilting at windmills for many decades.

Empirically, however, what about the likelihood of finding serial dependency in behavioral data from typical operant experiments? For the present paper, 24 graphs of experimental results were sampled from *JABA* and were reanalyzed using time-series procedures. Criteria for selecting this sample and the re-analyses are discussed later. For now, the evidence for serial dependency in these 24 experiments will be considered to show how common autocorrelation is in behavioral scores. Twenty of the 24 experiments (83%) had significant lag 1 autocorrelations, ranging from 0.40 to 0.93. Nine of the 20 significant autocorrelations were greater than 0.70. Clearly, then, serial dependency is a relatively common property of behavioral scores obtained in operant experiments.

Given that serial dependency is likely to occur in behavioral scores, what are the implications of this for making inferences about behavioral change due to an intervention? It was suggested above that operant researchers make inferences about behavioral change based on visual estimation of certain statistical properties of their data, *e.g.,* stability, variability, overlap, and number of scores. These statistical properties, and estimations of them either visually or computationally, are difficult to interpret unequivocally when applied to scores that are not independent of each other.

The difficulty with interpreting statistical estimates obtained from serially dependent behavioral scores can be explained via the following analogous situation. Suppose we wish to develop norms for teasing behavior in young children. Teasing rates per minute are obtained for each child in several families by direct observation in home settings. Our final sample of data includes teasing rates for each of 50 children, aged 3 to 6 yr from 25 different families. To determine a norm for teasing behavior in this age group, we could simply average the teasing rates over the 50 children. We certainly could do the arithmetic; that is, add the 50 teasing-rate scores and divide the total by 50, to get an average teasing rate.

The issue is whether or not the 50 summed teasing rates are independent of each other in the same way that 50 rate scores would be independent if they were obtained for 50 children, each of whom was from a different family, *i.e.,* if there were as many families as children in the sample. Because of the strong possibility that siblings show correlated teasing rates, the 50 scores in our example cannot be interpreted as independent. The mean score, therefore, will be biased in comparison to a mean obtained in the other case, where the 50 children were from 50 different families, and the correlation between sibling teasing rates simply could not exist. Not only will the mean be biased; other statistics will be as well. In particular, the variance in such a sample of correlated scores will tend to be less than in a sample of independent scores. Similarly, the variance or variability in serially de-

pendent time-series behavioral scores will tend to be less than in a series of serially independent scores.

The point is that we are accustomed to thinking about statistics as if they were always used with samples of independent scores, largely because statistical concepts were originally developed for these kinds of scores. When conventional statistical procedures are applied to serially dependent samples of scores, the assumptions underlying the statistics are violated, in particular the independence assumption. Our conventional graduate-school-learned statistical procedures and concepts were never meant to be used with serially dependent scores.

To summarize, then, serial dependency is evidenced by autocorrelations in an individual subject's behavioral scores obtained in reversal or multiple-baseline experiments. Serial dependency appears to be a common characteristic of behavioral scores for the individual subject, but one that is not well understood *vis-à-vis* its implications for either visual or statistical appraisal of changes in behavioral scores. Estimates of behavioral score properties like stabilities, variabilities, or averages, whether obtained visually or statistically, may be biased by serial dependency and if so, would not be as readily interpretable as if the estimates were obtained from independent scores.

Recommended Use of Time-Series Analysis in Operant Studies

The principal recommendation of this paper is that time-series analysis be used to supplement visual analysis of behavioral change in operant experiments. Used as a supplementary procedure, time-series analysis should confirm valid and appropriate inferences and conclusions derived from visual analysis. If visual and time-series analyses produce contradictory inferences or conclusions from the same set of data, then the visual analysis should be further scrutinized, since serial dependency may have misled the visual analysis.

This recommendation requires no further qualifications regarding the appropriateness of the method for behavioral data of the kind usually found in applied behavior analysis. Statistical purists, however, would be bothered by application of the time-series method to data that possess certain characteristics, *e.g.,* unequal numbers of scores in the different phases of an experiment, too few scores within or across the phases, or unequal time intervals between the scores in the series.[3] The reasons for concerns such as these are statistical in nature. The existence of such properties may bias parameter estimation in the time-series method, with the result that statistical significance tests will be conservative. This means that statistically significant findings will be less likely to occur when there are unequal or too few scores in the phases of an experiment.

It should be added that even visual analysis of operant studies could be biased as well when properties such as these obtain in an operant experiment. But in the applied and practical world of applied behavior analysis, it may not always be possible to obtain equal numbers of scores across phases or more than a few scores within certain phases. Hence, the operant researcher often has to make do with the available data. We argue, simply, if this is the case and the researcher is willing to be confident in visual analysis of possibly inadequate data, then the added support provided by time-series analysis for these visually derived inferences should be welcome. If confirmation of visual inferences is not obtained, the researchers should certainly entertain the possibilities that the visual analysis is inaccurate, and that the experimental data are simply not adequate to the task of inferring anything about behavioral change due to the experimental manipulation. The counter-suggestion that time-series analysis may be inaccurate or biased could also be entertained, but at least one likely cause of an inadequate visual analysis, *i.e.,*

[3]These purists would also be greatly bothered by simple visual analysis of time series, whether or not these properties of the scores were obtained.

serial dependency, is explicitly accommodated by time-series analysis.

Conceptual Description of the Time-Series Method

Three issues in the analysis of behavioral time-series scores should be of interest to users of reversal or multiple-baseline designs: (*a*) change *versus* no change in level, (*b*) zero *versus* nonzero trend, and (*c*) change *versus* no change in trend. Level is a formal term used to describe a location parameter in a serially dependent time series. The term level is reserved for autocorrelated data, while mean is used to describe the central tendency in uncorrelated data. If a serially dependent time series has been transformed to uncorrelated scores, then the level and mean are the same. But when we speak of the central tendency in autocorrelated time-series scores, we use the term level. Change in level refers to change at the interruption point, and is seen as discontinuity in the series from one phase to the next. Trend refers to any gradual upward or downward linear slope in the behavioral scores over time. Continuous trend may be evidenced throughout the entire series, perhaps running across all phases of the design. Trend characteristics of the data may vary from phase to phase of the reversal design. Thus, the third property is change *versus* no change in trend at each interruption point.

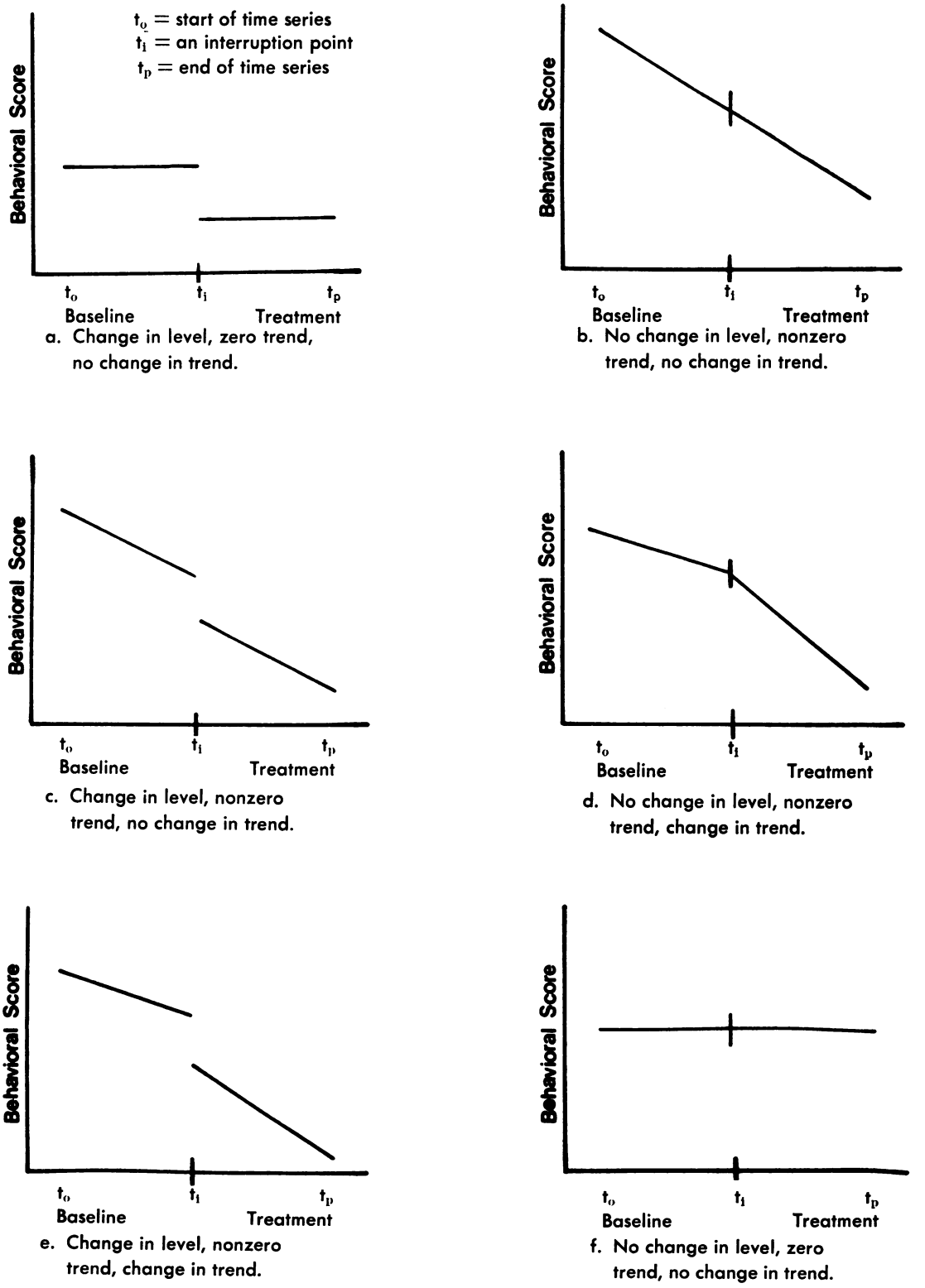
These three properties, and the two states of each, provide eight combinations of effects that could be of concern to operant researchers. But the two combinations involving both no overall trend and yet a change in trend are unlikely, so they will be ignored here. The remaining six combinations are described more fully below, and are illustrated graphically in Figure 1.[4] For simplification, only the baseline and initial treatment phase of a reversal design will be shown. Each graph shows a baseline and treatment phase, separated by an interruption point (t_i). The straight line "curves" are obviously idealized, given the usual variation of behavioral scores in operant research. Finally, the graphs have been drawn showing downward changes in the behavioral scores. The points to be made are equally germane for graphs showing upward changes.

a. Change in level, zero trend, no change in trend (Figure 1-a). This combination of effects illustrates a frequent hypothesis in operant research, change in behavioral scores from baseline to treatment. Change in level at an interruption point may be seen as an overall increase or decrease in the scores during the treatment period. The time-series procedure inspects the pattern of baseline scores and predicts what the scores in the treatment phase should be, given the baseline scores. If the scores in the treatment phase are statistically different from the estimated scores, the analysis estimates the magnitude of change and a probability value for the significance of the change.

b. No change in level, nonzero trend, no change in trend (Figure 1-b). The second kind of change that time-series analysis can detect involves any upward or downward linear slope in the behavioral scores over time. This kind of change, called trend, across all phases is particularly troublesome, since a treatment effect can be interpreted simply as a continuation of the trend first established during baseline, rather than a change in level at the interruption point due to treatment. For example, if treatment is designed to reduce the rate of deviant behavior, and baseline scores show a downward trend, then a reduction in deviant behavior during treatment might be misinterpreted as an effect of treatment, when in fact the observed reduction was nothing more than baseline trend continuing into the treatment phase. It is for this reason that operant methodologists (*e.g.,* Sidman, 1960) have cautioned investigators to establish stable (*i.e.,* zero trend) baselines before implementing experimental interventions.

c. Change in level, nonzero trend, no change in trend (Figure 1-c). It is precisely in such cases

[4]For additional kinds of changes in interrupted time series, see Glass, Willson, and Gottman (1975).



a. Change in level, zero trend, no change in trend.

b. No change in level, nonzero trend, no change in trend.

c. Change in level, nonzero trend, no change in trend.

d. No change in level, nonzero trend, change in trend.

e. Change in level, nonzero trend, change in trend.

f. No change in level, zero trend, no change in trend.

Fig. 1. Six illustrative treatment effects: combinations of level and trend changes detectable by time-series analysis.

of nonzero trend that time-series analysis is particularly useful. The method can detect changes in level even when nonzero trend exists in the scores. That is, if scores are gradually decreasing, and the treatment is powerful enough to produce the intended effects, time-series analysis may detect a significant change in level at the interruption point, taking into account the nonzero trend. Note the difference between Figure 1-b and Figure 1-c. Both show nonzero trend and no change in trend, but Figure 1-c shows a change in level at the interruption point while Figure 1-b does not. One obvious and important implication of these two illustrations is that investigators who use time-series analysis can be less concerned with establishing stable or zero-trend baseline scores than in the past. This is not to say that zero-trend baselines are unnecessary, but in some studies the cost of continuing baseline observations until zero trend is obtained may be prohibitive. In such studies, the data can often be analyzed by the time-series procedure to detect changes in level, even given the nonzero trend during baseline.

d. No change in level, nonzero trend, change in trend (Figure 1-d). The fourth kind of change that time-series analysis can detect involves changes in trend between baseline and treatment phases where there is no change in level at the interruption point. For example, a gradual downward change during baseline may be accelerated or slowed during the treatment phase. Detecting changes in trend essentially amounts to identifying changes in the rate of change in the scores. Change in trend might involve a reversal of the direction of the trend from baseline to treatment phases. That is, downward trend during baseline may change to upward trend during the treatment phase. Time-series analysis can detect such changes due to treatment where there is no change in level from baseline to treatment.

e. Change in level, nonzero trend, change in trend (Figure 1-e). The fifth combination of changes that time series can detect involves both level and trend changes. That is, treatment might change the level of scores at the interruption point, and also might change the rate of any upward or downward trend from baseline to treatment phases. When such changes in both trend and level parameters are obtained, visual inspection can be misleading, since it is difficult to accommodate change in trend subjectively when inspecting for change in level, or *vice versa.* But the time-series method can account for one of the two parameters when testing for significant change in the other.

f. No change in level, zero trend, no change in trend (Figure 1-f). The final case is, of course, what every operant researcher dreads—no changes in behavior. When the scores in both phases are highly variable, visual inspection for change in level can be misleading. Time-series analysis provides an appropriate method for testing the null hypothesis in instances of this kind.

What specifically does time-series analysis do with behavioral data? First, time-series analysis transforms the raw scores to uncorrelated or serially independent scores. Second, time-series analysis statistically compares the transformed scores from adjacent phases in the design. The transformation to serially independent scores is conceptually similar to other kinds of data transformations familiar to most social scientists (*e.g.,* arc-sin transformations of proportion scores, square-root transformations of low-rate frequency scores, logarithmic transformations of scores with similar means and standard deviations, or z-transformations of correlation coefficients). Each of these better-known transformations is used to change the properties of scores to meet the assumptions of parametric statistical tests. In the case of the time-series transformation, the resultant scores are freed of serial dependency, and therefore satisfy the assumption of uncorrelated error required by linear parametric methods, *e.g.,* analysis of variance.

The second thing that time-series analysis does is also conceptually uncomplicated. Readers familiar with the general linear model (Cohen, 1968; Overall and Spiegel, 1969; Walberg,

1971) will recall that the analysis of variance is essentially a special case of multiple regression analysis. The time-series method presented here uses the linear regression model to test for differences in level and trend. In the regression procedure, a dummy coding technique (Cohen, 1968; Kerlinger and Perlhauzer, 1973) is used to create variables whose values represent membership in ANOVA groups. The only difference between time-series and regression procedures lies in the value assigned to these dummy variables. In the regression technique, the usual values are either one or zero, reflecting group membership or nonmembership, respectively. In time-series analysis, comparable dummy coding is used to represent the baseline *versus* treatment "groups" of scores, but the values of the dummy variables incorporate a parameter that represents the amount of serial dependency in the raw data. If this parameter is zero, which would mean there is no serial dependency in the scores, then the time-series dummy codes reduce to precisely those used in the standard regression procedures for testing differences in means. For the present purposes, the point to remember is that the time-series method is procedurally identical to regression analysis for testing change in level, trend, and change in trend between baseline and treatment phases. However, in time-series analysis, the values assigned to the dummy variables are functions of the serial dependency in the scores, whereas in standard regression analysis, no dependency is reflected in the dummy codes.

Thus far, the presentation of time-series analysis has been as nontechnical as possible. Now would be the appropriate point to discuss the mathematical details of the time-series method, before reporting illustrative re-analyses of published studies where the time-series method was used to supplement the authors' original visual appraisals. But to keep this paper nontechnical, the reader interested in these details is referred to the literature cited earlier. In particular, we recommend starting with the book by Glass, Willson, and Gottman (1975).

Illustrative Time-Series Analyses

To demonstrate the utility of the time-series method in the analysis of operant data, examples were selected from the *Journal of Applied Behavior Analysis*. Experiments were chosen for reanalysis on the basis of the following criteria. First, experimental effects claimed by the authors and depicted in the graphs had to be sufficiently nonobvious to warrant some critical inspection. Second, to represent the variety of operant experiments, 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, particular attention was given to experiments where possible nonzero trend was apparent from visual inspection of the graphs. The examples discussed below cover five different operant research designs: (*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 examples involve combined data for a number of subjects, others involve individual subjects.

The first illustration is taken from a study of reinforcement principles in a military phychiatric ward. Experiment III from Boren and Colman (1970) was designed to increase attendance of soldiers at a unit meeting. The subjects had been operating in a token system where each soldier received 20 points (used to buy backup reinforcers) for attendance at 8:00 a.m. daily meetings. The investigators sought to increase attendance rate over that provided by the points alone, so applied a chaining contingency, whereby attendance at the meeting was required before additional points could be earned for satisfactory performance in other activities during the day. Fifteen days of attendance data were

used as a baseline or pretreatment period, during which the 20-point token system was in effect (Condition C). This phase was followed by 20 days of the chaining contingency (Condition D). The data are shown in Figure 2 (Figure 3 in Boren and Colman). Visual inspection of these data suggests two effects, first an increase in attendance due to the chaining treatment; and second, an upward trend in the scores, particularly during the chaining phase. Boren and Colman (1970) interpreted these data as follows:

> Within the first week, the participation increased to a median of 63%, compared to a median of 38% for the previous week. The attendance continued to rise until the median of the last week shown in Figure 4 [*sic*] was 87%. Since the data points in Figure 3 for the chaining condition overlap with only one data point for the previous condition, it is clear that the chaining technique increased attendance at the unit meeting with considerable reliability and without any increase in the points offered. [p. 33]

This interpretation suggests both changes in level of attendance due to the treatment, and upward trend in attendance during the chaining condition ("The attendance continued to rise. . ."). Time-series analysis was applied to the data to supplement these visually based interpretations. The regression analysis for change in level yielded a time series t of 2.84 ($df = 33$; $p < 0.01$), supporting the authors' interpretation of an increase in attendance during the chaining condition. But time-series tests for trend and change in trend were nonsignificant, contradicting the interpretation that attendance continued to rise during the chaining condition. Hence, while time-series analysis supports the conclusion of an abrupt increase in attendance due to the chaining procedure, evidence for a continued increase throughout the treatment was not obtained, contrary to what visual inspection of the data might suggest.

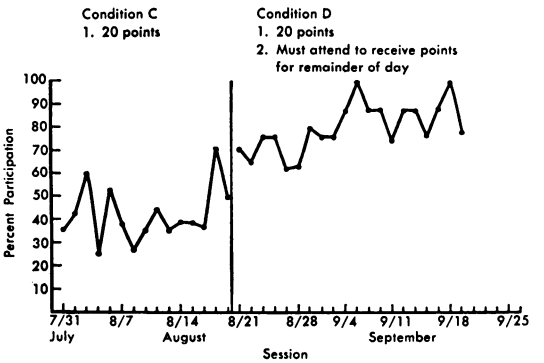


Fig. 2. Illustrative data from Experiment III in Boren and Colman (1970).

The next illustration uses data from Ingham and Andrews (1973), who used a traditional reversal design (ABAB) to assess the impact of token reinforcement and a penalty schedule on stuttering behavior of adults. An ABAB experimental design was used to make two comparisons between the token reward system (A), which was followed by a combined reward/penalty schedule (B). Speech therapists recorded the number of syllables spoken, the number of syllables stuttered, and the speaking time for each subject during 45-min rating sessions. The percentage of syllables stuttered and the rate of speech were computed for 21 sessions and are shown graphically in Figure 3 (Figure 1 in Ingham and Andrews). The authors were primarily concerned with evaluating the trend of the two dependent variables when the token penalty was removed from the overall system. This implies that their hypothesis was focused on an anticipated change in trend—not in level—between adjacent phases. The following conclusions were drawn:

> The penalty schedule combined with the reward schedule in the token system increased rate of reductions in frequency of stuttering and improved rate of speech, when compared with the reward schedule alone. [p. 233]

Ordinarily, such a statement would apply mainly to differences in mean levels. However, since the hypothesis was stated in terms of trend, time-

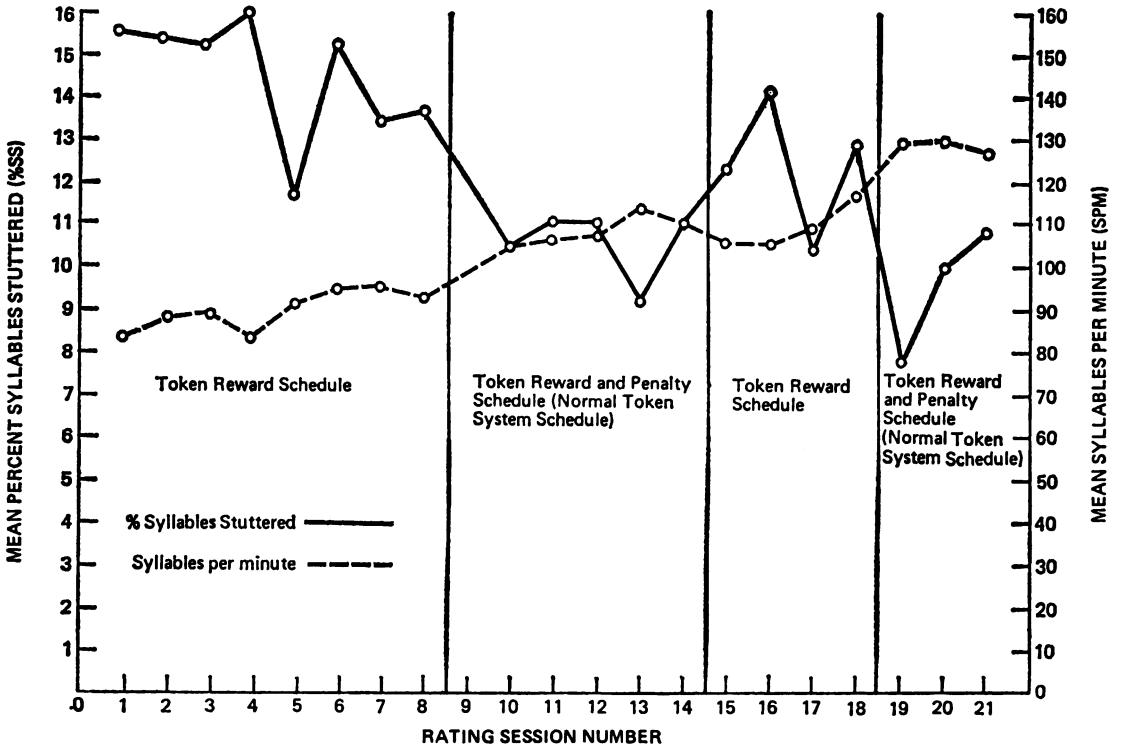


Fig. 3. Illustrative data from Ingham and Andrews (1973).

series tests for changes in trend were first conducted.

A significant nonzero trend was obtained for syllables per minute (dotted line) across all phases ($F = 15.42$; $p < 0.01$), but no changes in trend between adjacent phases were detected for either syllables per minute or per cent syllables stuttered (solid line). Therefore, the penalty component was not effective in altering the trend observed for the reward system alone.

When the time-series procedure was used to test for changes in level, only the final manipulation proved sufficiently powerful to impart clear improvement in syllables per minute ($t = 2.18$; $df = 5$; $p < 0.05$). No changes in the level of this variable were observed between other phases. Also, no changes in level for per cent syllables stuttered were obtained between adjacent phases.

Note that the one significant change in level obtained involved only three data points in the last phase. Even with this small number of points, the time-series method produced significant findings. This does not, however, mean that only three data points are recommended for experiments analyzed by time-series methods. In fact, in this example, it is not clear that the reward/penalty schedule was the sole reason for the significant change in level from Phase 3 to Phase 4. The change from Phase 1 to 2 was not significant. The significant change from Phase 3 to 4 could have been due to a cumulative effect of the contingencies operating across Phases 1 through 3. Of course, further experimentation would be needed to support this *post hoc* interpretation.

The third illustration of the time-series method uses Experiment I in Phillips *et al.* (1971), which was designed to increase promptness at evening meals at Achievement Place. During a 35-day baseline phase, elapsed minutes were recorded between the ringing of the dinner bell and the time at which the last of the four boys in the study sat down at the dinner table.

Next, during each of 20 days of the "points" condition, each boy lost 100 points for each minute he was late. This first treatment condition (B) was followed by a second condition (C) in which threats of loss of points were announced periodically during the 19 days of this "Threats No Points" condition, but no points were deducted. The final phase of this experiment (16 days) was a return to the points condition (B), where, again, each boy lost 100 points for each minute late. The graph of these data is shown in Figure 4 (Figure 1 in Phillips *et al.*). The authors' interpretation of these data was:

> When the 100-point fine was made contingent on each minute late, the boys were more prompt and, by the end of the Points condition, all boys were seated at the table less than 60 sec after the dinner bell rang. Under the Threats No Points condition, the behavior reversed to about 10 min late; during the final Points condition the boys were again prompt in coming to dinner. These data indicate that point losses were effective in producing punctual behavior at dinner time. [p. 47]

This interpretation suggests that the level of "minutes late" was reduced by the Points condition (B), was increased by the Threats No Points condition (C), and was again reduced by re-institution of the Points condition (B). No suggestion of trend or change in trend is evident in either the authors' interpretation or the graph of these data.

Time-series analyses of these data produced a significant F for changes in level ($F = 18.11$; $df = 3, 81$; $p < 0.001$), but nonsignificant F-ratios for trend and change in trend. These supplementary time-series findings illustrate the ap-

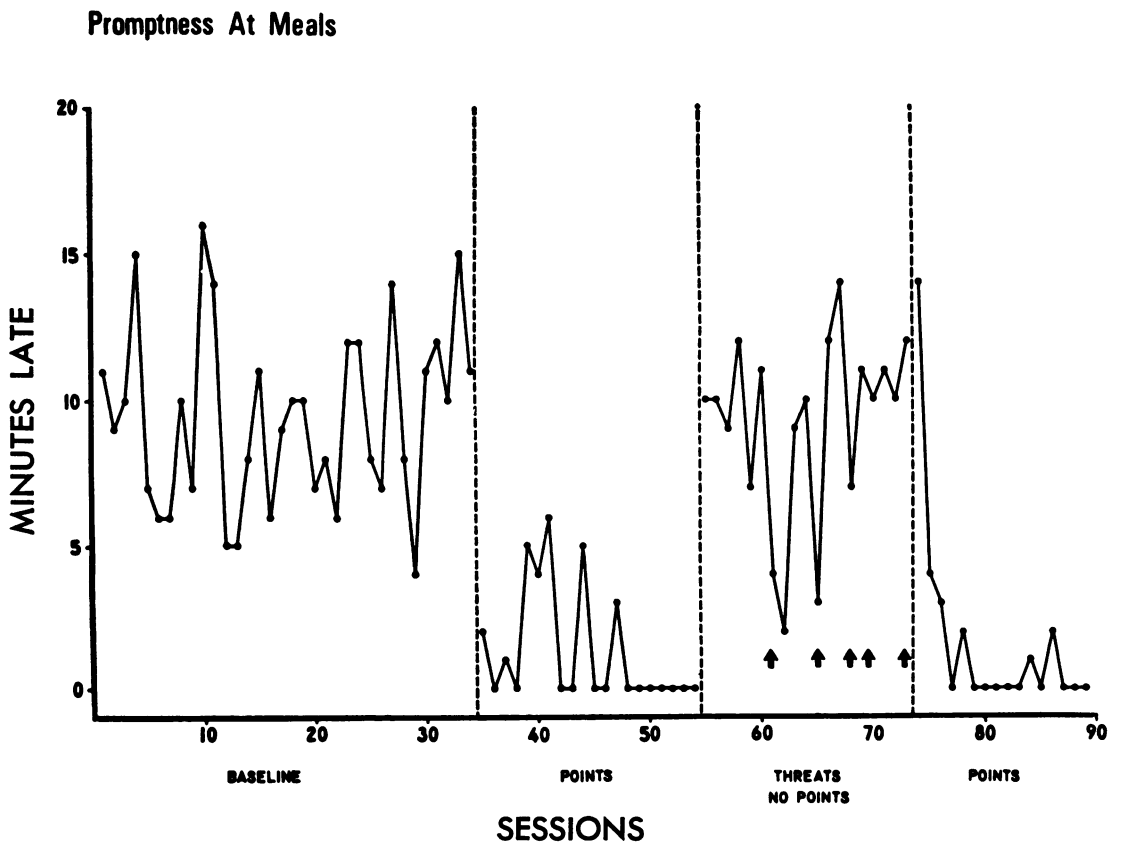


Fig. 4. Illustrative data from Experiment I in Phillips, Phillips, Fixsen, and Wolf (1971).

plication of the time-series method to a multiple component operant study and tend to confirm the authors' original interpretations, although comparisons between adjacent pairs of phases would be required to duplicate the authors' conclusions precisely.

The fourth example (Baer *et al.*, 1973) used a multiple-baseline design to investigate the effects of differential reinforcement on the compliance of three negativistic preschool children. Baer *et al.* modified an ongoing token system to include teacher-delivered reinforcement for completion of specifically suggested tasks. In addition, timeout for noncompliance was applied in subsequent experimental phases for two subjects. During 60 days of observation, compliance to the teacher's formal invitations to tasks were recorded. A graphic representation of these data is presented in Figure 5 (Figure 1 in Baer *et al.*). Visual examination suggests that changes in level were obtained for all subjects when reinforced for requested completion (Phase B). The authors concluded that:

> For each child, this technique resulted in clear and useful increases in compliance. . . . In the case of two children whose compliance was not maximized by differential reinforcement alone, further increases in compliance were produced by combining a 1-min timeout for noncompliance with the differential reinforcement procedure. . . . [Timeout] was more effective than differential reinforcement alone. [pp. 289, 297]

Here, the authors claim a second change in level for the two children subjected to the timeout contingency (Phase BC).

Three separate time-series analyses were performed, one for each subject. The regression analysis for change in level yielded a t of 2.08 ($df = 28$; $p < 0.05$) between Phases A and B for Hannah, and a similar result for Frankie ($t = 7.35$; $df = 58$; $p < 0.01$). Thus, the supplementary time-series analysis supported the authors' contention that the reinforcement for requested completion was effective in raising compliance above baseline levels. However, this conclusion is not supported for Charlotte ($t = 1.63$; $df = 28$; n.s.). Further, the claim that timeout (Phase BC) increased compliance was not substantiated for either Charlotte or Hannah. All t's for changes in level between adjacent Phases 2 through 5 for Charlotte and Hannah were nonsignificant.

The last example of time-series analysis uses data from a reversal component study conducted by Wincze *et al.* (1972). In an effort to reduce delusional verbal behavior of hospitalized schizophrenics, several token reinforcement procedures were implemented following a therapist's feedback condition. These procedures included differential application of tokens in therapy sessions and on the ward, as well as a bonus system for remaining below a percentage criterion level of delusional statements. The verbal behavior of 10 subjects was monitored for 21 to 63 days, during which various contingencies were operating within a counterbalanced design. Only some of the findings for Subject 7 will be discussed, as an illustrative example.

The percentage of delusional talk for Subject 7 is shown graphically in Figure 6 (Figure 7 in Wincze *et al.*). The present discussion is confined to delusional behavior in therapy sessions (the open circles). Conclusions of the authors are as follows:

> Token reinforcement reduced the percentage delusional talk in therapist sessions (phases 2, 6, and 7). . . . Feedback applied in therapist sessions (phase 4) slightly reduced the percentage delusional talk. [p. 254]

All eight phases of the experiment were included in the time-series analysis. Comparing the level in each phase with the level of the preceding phase produced only one statistically significant change in level, between Phases 6 and 7 ($t = 2.36$; $df = 12$; $p < 0.05$). Thus, only

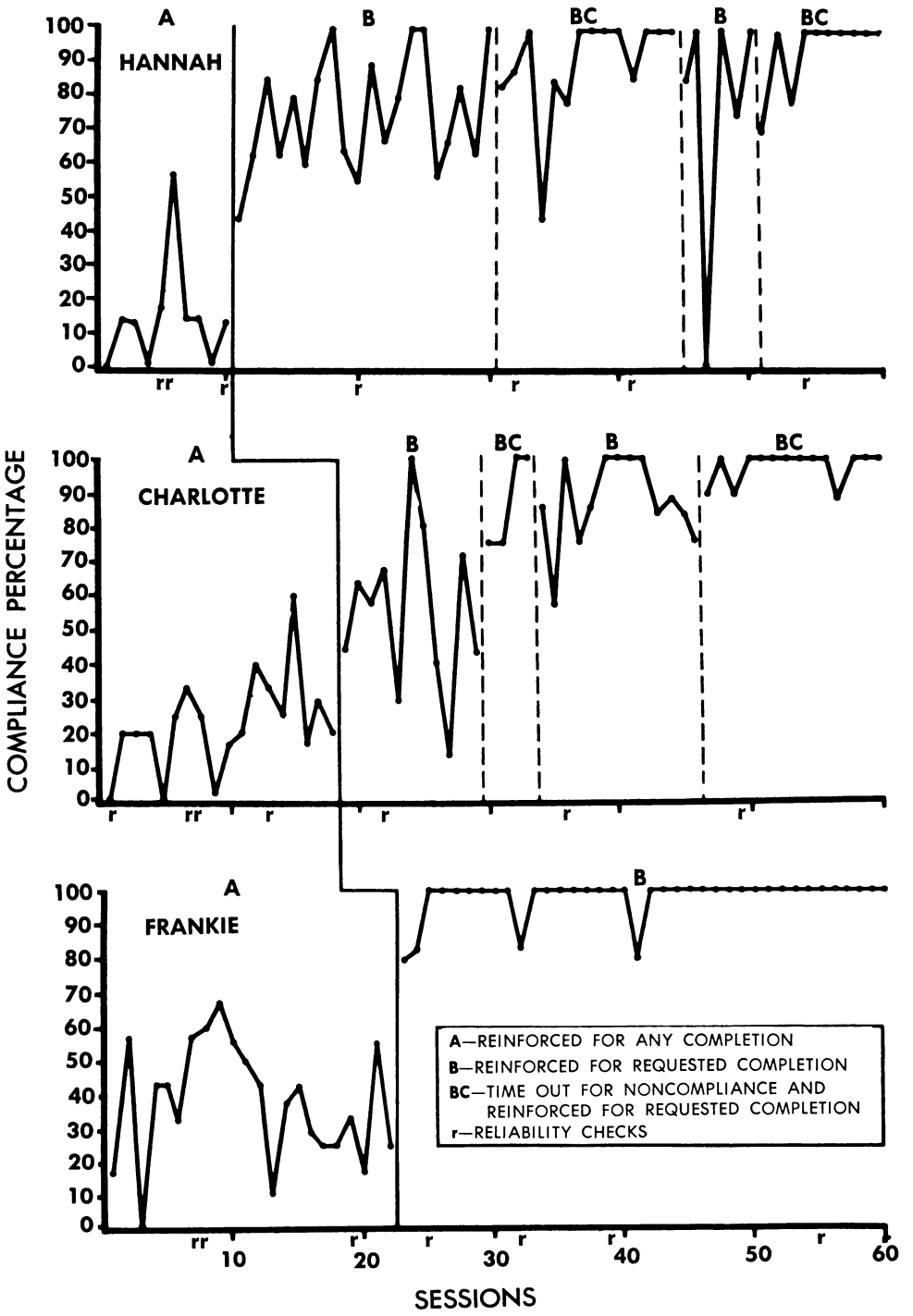


Fig. 5. Illustrative data from Baer, Rowbury, and Baer (1973).

one of the four phase changes claimed by the authors was supported by the time-series analysis.

But, note that there are visually apparent trends in several of the phases, particularly Phases 1, 3, 4, and 8. It is the presence of these

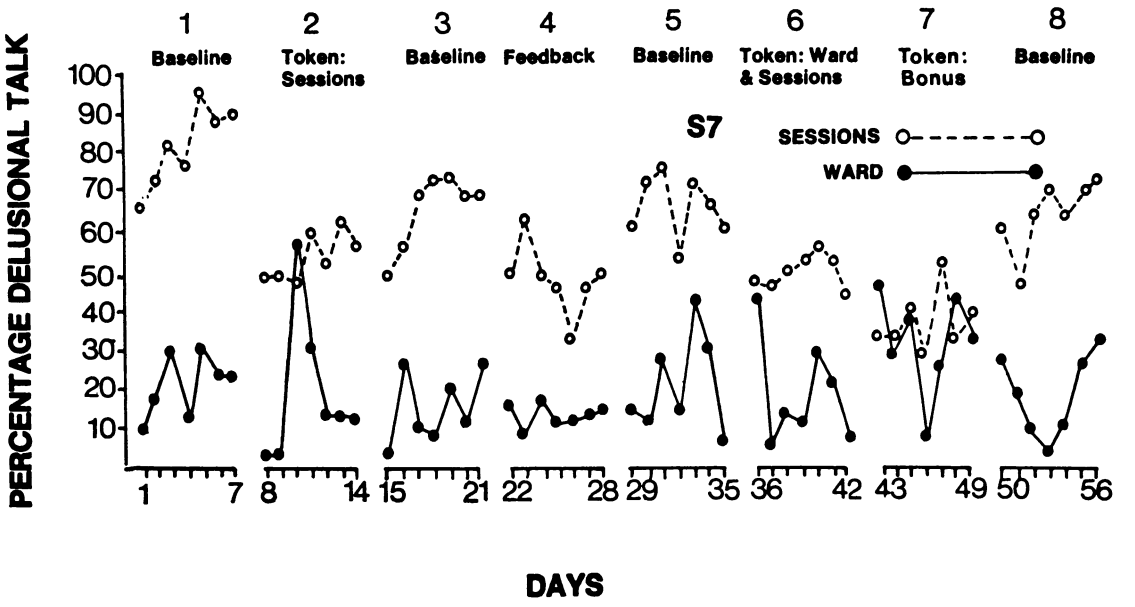


Fig. 6. Illustrative data from Wincze, Leitenberg, and Agras (1972).

trends that makes claims for changes in level equivocal, and which produced the nonsignificant level changes. Time-series analysis does allow study of changes in trend, however, and is particularly useful when stable (*i.e.*, zero trend) baselines are not obtained, as in this study. Hence, the time-series procedure was used to test for changes in trend, particularly between those phases for which the authors claimed level changes, but for which the *t*'s were nonsignificant. Significant differences in trend were obtained between baseline Phase 3 and feedback Phase 4 ($t = 3.18$; $df = 12$; $p < 0.01$), between feedback Phase 4 and baseline Phase 5 ($t = 2.67$; $df = 12$; $p < 0.05$), and between baseline Phase 5 and Phase 6 ($t = 2.26$; $df = 12$; $p < 0.05$).

These results show that the effects of intervention were to change the trend characteristics (upward or downward) from one phase to the next, but as already noted, changes in level were not obtained, given the trends extant in the data (except between Phases 6 and 7, where little if any trend occurred). Thus, changes in trend were inappropriately interpreted as changes in level, based on the authors' visual analysis of their data. Now, it still could be claimed that the interventions were effective, but the nature of the effects was to change the trends, not the levels. One could argue that the trend changes amount to the same thing as the claimed, but not supported, level changes. However, this interpretation would be supported unequivocally only if the phases had been lengthened to allow stable (*i.e.*, zero trend) scores to emerge. When trend is evident and relatively few data points are used, claims for level changes will often be equivocal.

DISCUSSION AND CONCLUSIONS

Other illustrations of the time-series method could be presented, using data taken from *JABA,* but they would not add appreciably to the present discussion. From the previous discussion and the examples shown above, the following conclusions seem justified.

First, the examples covered the variety of operant experiments reported in *JABA,* including reversal designs and multiple-baseline designs with multiple treatment or intervention phases. Data for individual subjects, as well as aggregate data over a set of subjects, were analyzed. Designs with unequal numbers of scores in the various phases of the experiment were used, and

the range of number of data points within phases was considerable, *i.e.*, from only three to as many as 34. Statistically significant findings were obtained for changes in both level and trend, and these findings were obtained between phases with the smallest, as well as the greatest, number of scores. In short, the time-series method illustrated here was satisfactorily applied to typical applied operant designs, which incorporated a great variety of score and design properties. Thus, it seems fair to conclude that the time-series methods used are satisfactory supplements to visual methods for analysis of operant data.

Second, casting time-series analysis as a procedure to supplement visual analysis, the illustrations show that in many instances, the authors' visually based conclusions were supported, in other cases they were not, and in still others the time-series analysis revealed findings that had not been discussed by the original experimenters. All three kinds of supplementary information provided by time-series analysis are useful. It is rewarding to have one's visual impressions supported by statistical analysis. It is humbling and/or educational to have other impressions not supported. And it is clearly beneficial to have unseen changes in the data detected by a supplementary method of analysis. It is difficult to see how operant researchers can lose in the application of time-series analysis to their data. Probably no statistical method will ever replace human judgment (Michael, 1974), but as a supplementary tool, time-series analysis deserves a place in the operant methodólogists' armamentorium.

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.

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.

Campbell, D. T. and Stanley, J. C. *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally, 1970.

Cohen, J. Multiple regression as a general data-analytic system. *Psychological Bulletin,* 1968, **70,** 426-443.

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.

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

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.

Kerlinger, F. M. and Perlhauzer, E. J. *Multiple regression in behavioral research.* New York: Holt, Rinehart & Winston, 1973.

Michael, J. Statistical inference for individual organism research: some reactions to a suggestion by Gentile, Roden, and Klein. *Journal of Applied Behavior Analysis,* 1974, **7,** 627-628.

Overall, J. E. and Spiegel, D. K. Concerning least squares analysis of experimental data. *Psychological Bulletin,* 1969, **72,** 311-322.

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

Risley, T. R. and Wolf, M. M. Strategies for analyzing behavioral change over time. In J. Nesselroade and H. Reese (Eds.), *Life-span developmental psychology: methodological issues.* New York: Academic Press, 1972.

Schnelle, J. F. and Lee, J. F. A quasi-experimental retrospective evaluation of a prison policy change. *Journal of Applied Behavior Analysis,* 1974, **7,** 483-496.

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

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

Walberg, H. J. Generalized regression models in educational research. *American Educational Research Journal,* 1971, **8,** 71-91.

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.

Received 12 June 1974.
(Final acceptance 22 April 1976.)