JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1980, 13, 543-559 NUMBER 4 (WINTER 1980)

INTERRUPTED TIME-SERIES ANALYSIS AND ITS APPLICATION TO BEHAVIORAL DATA

DONALD P. HARTMANN, JOHN M. GOTTMAN, RICHARD R. JONES, WILLIAM GARDNER, ALAN E. KAZDIN, AND RUSSELL S. VAUGHT

UNIVERSITY OF UTAH; UNIVERSITY OF ILLINOIS; EVALUATION RESEARCH GROUP, EUGENE, OREGON; UNIVERSITY OF UTAH; THE PENNSYLVANIA STATE UNIVERSITY; AND THE STATE UNIVERSITY OF NEW YORK AT BINGHAMTON

This paper uses a question-and-answer format to present the technical aspects of interrupted time-series analysis (ITSA). Topics include the potential relevance of ITSA to behavioral researchers, serial dependency, time-series models, tests of significance, and sources of ITSA information.

DESCRIPTORS: data analysis, methodology, serial dependency, statistics, time-series analysis

Various techniques have been proposed to aid in judging the significance of change in individual subject research. These decision aids include visual analysis of graphic displays and rule-of-thumb, as well as formal inferential procedures applied to descriptive statistics. Perhaps the most suitable statistical procedure for analyzing individual subject data is interrupted time-series analysis (ITSA). Unfortunately, however, ITSA is technically complex, and at present there are few papers that bridge the gap between the elementary description by Jones, Vaught, and Weinrott (1977) and the more complex mathematical presentation by Glass, Willson, and Gottman (1975). This paper is intended to help bridge that gap and hence ease the pains of entry into the technical ITSA literature.

Portions of this paper were supported by the following research grants: HDMH 06914 from the National Institutes of Health, United States Public Health Services to Donna M. Gelfand and Donald P. Hartmann; NIMH Research Scientist Development Award 1K02MH00257 to John M. Gottman; and R01 MH 31018 from the Center for Studies of Crime and Delinquency, NIMH, United States Public Health Services to Richard R. Jones. Reprints may be obtained from Donald P. Hartmann, Department of Psychology, University of Utah, Salt Lake City, Utah 84112.

The reader should understand that he or she will not be able to perform an ITSA as a result of reading our paper. It is not intended as a cookbook for performing an ITSA. It is instead a description of the novel aspects of ITSA written in reasonably ordinary language. We have written the material that follows in a question-and-answer format. This format allows the reader the options of obtaining an answer to a specific question by turning to the relevant portion of the paper or obtaining a general understanding of ITSA and related issues by reading the paper straight through.

Q: What is interrupted time-series analysis?

A: Interrupted time-series analysis is a statistical method for analyzing temporally ordered scores to determine if an experimental manipulation, a clinical intervention, or even a serendipitous intrusion, has produced a reliable change in the scores. Unlike other decision aids, such as visual analysis or analysis of variance, ITSA accommodates serial dependency, a common property of single organism behavioral scores. Serial dependency violates assumptions underlying traditional statistical models, such as the analysis of variance (Glass et al., 1975; Gottman & Glass, 1978),

and appears to hinder the use of visual analysis as well (Jones, Weinrott, & Vaught, 1978).

Q: Why might a behavioral researcher be interested in ITSA?
A: The usual visual method for the analysis of experimental effects may be unreliable in two important cases: when the experimental effect is small or otherwise difficult to detect, and when the observations are serially dependent. If neither of these conditions exists in a given study, then visual analysis of graphic representation of data will probably produce reliable inferences about the effects of a manipulation (see Parsonson & Baer, 1978, and Tukey, 1977, for a thorough discussion of graphic analysis).

In the first case, when an experimental effect is small or difficult to detect, one or more of the following tends to occur: (a) the results of visual analysis will be less reliable when the effects of manipulations are small; (b) baseline trends or cycles are hard to separate from the behavior changes caused by the manipulation; and (c) the eye has trouble distinguishing real behavior change from random behavioral fluctuations when scores are highly variable. Of course, applied analysts may not be interested in any statistical method for analyzing data (Michael, 1974; Baer, 1977; Parsonson & Baer, 1978; for other views, see Jones et al., 1977; Gottmann & Glass, 1978; and Kazdin's summary, 1976). Opponents of statistical analysis have proposed several alternative remedies for the data problems described above. Some would argue that small behavior changes are not worthwhile, even if shown to be statistically reliable (Baer, 1977), that unstable baselines should be discarded or continued until trends or cycles dissipate (Sidman, 1960), and that highly variable scores should be aggregated to reduce variability or the study should be rerun under more controlled (or controlling) conditions. But when these options are viewed as undesirable or impracticable, then ITSA should be a useful supplement to visual analysis (Jones et al., 1977).

The second problem, *serial dependency* among the observations, is a more subtle and, perhaps, more pervasive problem. Serial dependency refers to the fact that most time series—that is, temporally ordered behavioral scores for a single unit such as a subject, classroom, or family—do not consist of statistically independent observations (Glass et al., 1975). With serially dependent data, the performance of the unit at a given point in time can be predicted from its performance at one or more earlier points in time. When scores are serially dependent, visual analysis will agree with statistical analysis (ITSA) less often than when scores are not serially dependent (Jones et al., 1978). So visual analysis tends to produce less reliable and, therefore, less valid inferences from applied behavioral studies when this condition exists in the data than when it does not. And serial dependency is common in time-series data sets. Jones et al. (1977) reported finding serially correlated data in 83% of nonrandomly selected graphs from the *Journal of Applied Behavior Analysis,* whereas Kennedy (Note 1) reported finding only 29% of the graphs he analyzed from the same journal to have significant serial correlations. Although it is not clear how to account for these discrepancies, it is clear that a substantial number of published individual subject data sets are serially dependent.

Q: What are time-series data?
A: Time-series data are observations on some variable gathered at regular intervals. The observations may be obtained on a single subject or on an aggregate of subjects considered as a functional unit such as a classroom of children or a married couple. The set of potential observations is called the time-series *process.* Any real set of time-series data is properly called a *realization* of a process. Just as in ordinary statistics, where a set of data is referred to as a sample from a population, in time-series analysis a set of time-ordered data is referred to as a realization from a particular time-series process. For example, if we observed a child's thumb-sucking for 30 days, we would obtain

slightly different realizations depending on such factors as when we began observing and when we observed.

Q: What is a time-series model?
A: A time-series model is an attempt to describe mathematically a naturally occurring time-series process. The time-series data at hand (a realization of the process) are used to estimate the process. In ordinary statistics, we model our observations by splitting them into two components: the experimental effects component and the error component. Similarly, a complete time-series model has two components: *the deterministic component,* which reflects effects that are consistent across some period of time, and the *stochastic component,* which reflects recent error or noise. The deterministic component may include the mean of the process, an effect that we infer to be constant throughout the time of the process, and treatment effects, which we infer to be limited to particular experimental phases. The deterministic component of a time-series model exactly parallels the experimental effects component of ordinary statistical models such as the model for the independent group analysis of variance. However, the error component of ordinary analysis of variance and the component of time-series analysis are not entirely analogous. A time-series model subdivides the stochastic component into a part that reflects a random event occurring at the time of the observation and a part that reflects the effect of previous random event(s) on the observation. This latter systematic portion of the stochastic component is responsible for the serial dependency in the time series.

Q: What is serial dependency?
A: Serial dependency is the property of predictability among the stochastic or error components of the time-series observations that invalidates traditional statistical assumptions. When we use conventional statistics, we must assume that the error components are independent, and we take great pains to contrive situations in which this assumption will be valid. But when we intensively study the behavior of an individual unit, it is clear that the successive observations cannot be easily isolated and made independent.

Consider the following illustration: A male subject generates data during a weight control program through the daily observation of his own scale. Let us analyze the components of these observations. The weight of bone and vital organs will be constant throughout the realization. There may be an effect due to treatment or its absence during the different phases of the experiment. These are the deterministic components. There will also be a stochastic component of random, daily fluctuations around the expected levels of the deterministic effects. Suppose that on day one of the program the man's weight takes a random bounce above baseline level. When the man observes his weight in the evening, the observation of this deviation shocks him into abstinence during the next day. The observation on day one has had a reactive effect. On day two, the stochastic deviation from baseline includes the random fluctuation occurring on day two, plus the random event of day one times a negative factor reflecting the man's abstinence. This effect of the previous day's fluctuation is the systematic stochastic component. If the net effect of both of the stochastic components of day two's observation is a deviation below baseline, we might predict that the systematic component of day three's error component will tend to push the weight back up, and so on. Note that this predictability or dependency has nothing to do with the observations' absolute location in time, as the relations between stochastic components do not depend upon whether the experiment is in the baseline or treatment phase. The serial dependency between two observations is strictly a function of their position with respect to each other, of their adjacency in this example.

Q: How is serial dependency assessed?
A: Serial dependency is assessed by calculating

the autocorrelations between observations separated by different time intervals or lags in the series. A lag-1 autocorrelation is computed by pairing the initial with the second observation, the second with the third observation, and so on until the second from the last is paired with the last observation.

The lag-2 autocorrelation is calculated by pairing scores that are two intervals apart. For example, the initial is paired with the third score and the second with the fourth score. If the total number of observations is symbolized by n, we ordinarily compute $n/4$ lag correlations (Box & Jenkins, 1970, chapter 6), because as the lag increases there are fewer and fewer pairs of observations contributing to the correlation. For example, with $n = 72$, r_{18} has 54 pairs of scores contributing to the correlation.

In order for the autocorrelations to reflect the systematic component of the stochastic or error process, it is important that the observations be obtained at regular (equal) intervals. Observations conducted at irregular, or variable intervals are likely to disguise or alter the pattern of autocorrelations that would otherwise be obtained. This point can be illustrated with a simple example. The degree of wakefulness of many persons is predictable across 24-h cycles. Hourly observations of these persons' degrees of wakefulness would yield substantial autocorrelations at lag 24. If, however, observations were obtained at irregular intervals, the lag-24 autocorrelation might pair an observation on Monday at 4:00 a.m. with another observation on Wednesday at 1:00 p.m., rather than Tuesday at 4:00 a.m. Thus, the irregular observations would disguise the regularity in the wakefulness series.

The general formula for the lag-k autocorrelation or serial correlation, abbreviated r_k, is

$$r_k = \sum_{i=1}^{N-k} (Z_i - \bar{Z})(Z_{i+k} - Z) / \sum_{i=1}^{N} (Z_i - \bar{Z})^2,$$

where N is the total number of observations in the series, Z_i is the value of the observation at time period i, $\bar{Z}$ is the mean of the series, and k is the number of lags.

A substantial lag-1 autocorrelation is sufficient cause for concern about serial correlation, but it is not a necessary cause. The researcher should keep in mind that some cyclic phenomena can cause powerful serial dependencies at longer lags with or without a lag-1 correlation. "Seasonal" effects, such as might be found in monthly observations, are unlikely to be found in the relatively short series that are of interest to applied behavior analysts. (See McCain and McCleary, 1979, p. 261-273 for discussion of seasonal effects in time series.) Hourly measurements collected over many days could also reveal important dependencies at lag-24.

Q: How does serial dependency affect statistical tests?

A: Because serial dependency concerns only the stochastic component of the observation, it is not surprising to find that it does not bias our estimates of the deterministic parameters of the process, such as the mean. But it does bias estimates of the error variance and hence all conventional tests of significance. Negative autocorrelation, as was present in the weight example above, reduces the error term and hence gives traditional tests a conservative bias. Positive autocorrelation, probably a more common situation, increases the error term and creates a liberal bias when ordinary hypothesis testing procedures are used (Hibbs, 1974; also, see Scheffé, 1959, ch. 10). That is, far too many interventions are found to be statistically significant when no real effect exists. Because visual analyses also (and appropriately) take account of the apparent variability in the data to estimate the strength of an experimental effect, they too are biased in the presence of autocorrelation (Jones et al., 1978). Although Jones et al. found that serial dependency affects the reliability and accuracy of visual analysis, it is not yet known whether the effects of serial dependency on visual analysis parallel the effects

of serial dependency on conventional statistical tests.

Q: How does ITSA accommodate serial dependency?
A: A major goal of ITSA is to model the structure of the stochastic components of the time-series observations. Once a model is fitted to the stochastic component of the data, the systematic part of the error can be subtracted from each observation. The resulting scores are called residuals, and they contain no serial dependency. The residual scores meet the assumption of independence underlying ordinary statistical procedures. Techniques like t tests can then be applied to assess changes in behavior from one phase to another.

HOW TO CONDUCT AN INTERRUPTED TIME SERIES ANALYSIS

Q: What kinds of models are ordinarily fitted to the stochastic component of time-series data?
A: The most common models fit to the error or stochastic component of time-series data are the *A*uto*R*egressive *I*ntegrated *M*oving *A*verage (ARIMA) models. These models are described by Box and Jenkins (1976) and by Glass et al. (1975). *Autoregressive* and *moving average* are the two elementary models for the structure of the stochastic component of the time-series process.

Q: What are the autoregressive and moving average models?
A: The autoregressive (*AR*) and moving average (*MA*) models are the two forms of dependence that can be exhibited in the stochastic components of time-series observations. Processes are said to contain components which are "autoregressive *order-p*" [AR(p)] or "moving average *order-q*" [MA(q)]. The order of these processes denotes the number of prior observations that are included as terms in the systematic portion of the observation's stochastic component. For instance, the AR(1) model expresses each observation as a function of a random event and the previous (or lag-1) observation in the series. The AR(2) model expresses each observation as a function of a random event and the two previous observations.

Before further developing AR and MA models, we will examine a time series in which the data are serially *independent*. Panel A of Figure 1 is a simulated time series consisting of 100 independent samples from a standard normal distribution (generated by a computer). This time series has neither autoregressive nor moving average components and is commonly known as standard normal *white noise*. When we state that the realization in Panel A is standard normal white noise, we claim that if we carried the process out to a great length and drew a frequency distribution of the many values obtained, our graph would match the normal curve. But in any finite realization, we can only obtain an approximate match.

We will use Z_i to designate our observation

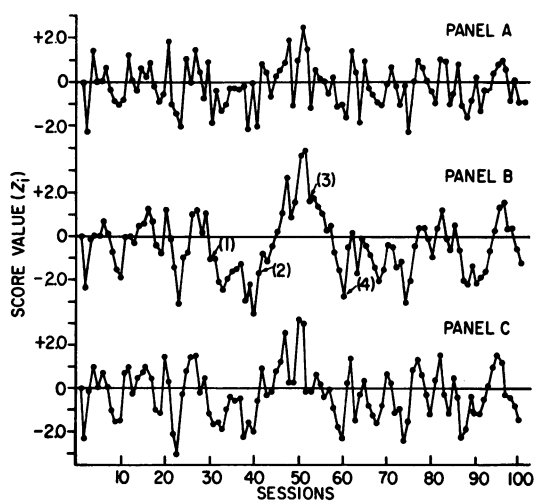

Fig. 1. Plots of (A) a white noise process, (B) the same white noise observations transformed by an autoregressive process, and (C) the white noise observations transformed by a moving average process. Arrow (1) in Panel B is the first observation in baseline of a hypothetical A-B-A experiment, Arrow (2) is the first point of treatment, Arrow (3) is the first point of the second baseline, and Arrow (4) is the last data point of the experiment.

of the process at time i. A_i is an independent sample from the standard normal distribution occurring at time i. So we can model the white noise process quite simply with the equation

$$Z_i = A_i.$$

These uncorrelated stochastic components exemplify the necessary conditions for valid conventional statistical tests.

Panel B of Figure 1 represents a graph of the same normal samples as were exhibited in Panel A, but made serially dependent through a transformation to an AR(1) process. In an AR(1) process, each observation Z_i again includes A_i as its random component. But Z_i also includes a systematic component: the parameter ϕ times the previous observation Z_{i-1}. That is,

$$Z_i = A_i + \phi(Z_{i-1}),$$

where ϕ (phi) is a number between 1 and -1. The larger the absolute value of ϕ, the more powerful the effect of Z_{i-1} upon Z_i, and consequently, the stronger the serial dependency within the data set. (In an AR(1) model, ϕ functions exactly as does r, the ordinary product moment correlation, in a standardized bivariate regression equation.) In Panel B we have chosen a substantial value of $\phi = 0.7$, in order to illustrate the effect of serial dependency.

Suppose that a researcher had begun observing this realization at session 30, marked with arrow 1, in Panel B. After 11 observations have shown a 'declining baseline,' the experimenter introduced a treatment at session 41, marked by arrow 2. Upon introduction of treatment, the slope of the graph turned abruptly and dramatically upward. After eleven 'treatment' observations, the researcher instituted a reversal phase at session 52, marked by arrow 3. Nine observations were recorded, and data collection ceased at session 60, marked by arrow 4. It is easy to see how one might be tempted to infer a deterministic structure within this time-series realization. But, in fact, the sustained negative and positive passages of the realization in Panel B are simply reflections of minor sampling fluctuations in the corresponding segments of the white noise realization shown in Panel A, "amplified" by the autoregressive process. Intuitively, the effect of the positive ϕ value in the Panel B series is to give the process "inertia"—a tendency to retain the effects of previous observations—and hence to exaggerate short-run variations within the white noise realization. A negative ϕ value, however, would give the opposite effect; that is, the series would oscillate rapidly around the mean.

Panel C illustrates the transformation of the Panel A observations into an MA(1) process. Once again, each observation Z_i includes the random component A_i. Now, however, the systematic component is θ (theta) times A_{i-1}, the previous *random event*. The equation for the MA(1) model is:

$$Z_i = A_i + \theta(A_{i-1}).$$

In Panel C, $\theta = .7$. The effect of the MA(1) process is once again to smooth the graph and give it "inertia." But here the effect is less extreme than that of the AR(1) process. That is, Z_{i-1} has more impact upon Z_i in the AR(1) model than in the MA(1) model. This is because only the random component A_{i-1} of observation Z_{i-1}, not the entire observation, appears in the MA(1) equation.

Q: What is the ARIMA (p,d,q) model?
A: Simple AR(p) and MA(q) processes are found in time-series realizations, but we also find processes that include components of both types. The autoregressive-integrated-moving average model therefore serves as a general model that includes the elementary processes as special cases.

ARIMA models have three parameters—usually symbolized p, d, and q—that must be estimated from the data. These three parameters completely describe the stochastic or error component of a time series. The parameter p indicates the autoregressive *order* of the model. A pure autoregressive-order 1 process, such as the one in Panel B of Figure 1, would be writ-

ten ARIMA (1,0,0). The q parameter refers to the moving average *order* of an ARIMA model. A pure moving average-order 1 process (Panel C of Figure 1) would be written ARIMA (0,0,1). An ARIMA process containing both an autoregressive and a moving average component would be represented by the following equation:

$$Z_i = A_i + \phi(Z_{i-1}) + \theta(A_{i-1}).$$

We would write this process as ARIMA (1,0,1).

The middle parameter in the ARIMA model, d, refers to the *order of differencing* that may be required in order for the series to meet a critical assumption of interrupted time-series analysis, called the *weak stationarity assumption.* Stationarity requires that the structure and parameters of the time-series process do not change as a function of time. Imagine that the time series has been separated into several different chunks. In practice, weak stationarity means that we must assume that the mean and variance and the autocorrelations are the same for each chunk of the series. Weak stationarity is important because whenever we do ITSA we are trying to say something about the future from the past, and we cannot do that unless we assume that some function of the time series is not changing. We must assume that some kind of regularity or stability with time exists or prediction is impossible.

Time series that are heterogeneous with respect to their mean, variance, or autocorrelations are called *nonstationary.* The stationarity assumption is violated, for example, by series that have secular trends, that is, by series that change in level by drifting up or down. Many time series in the behavioral sciences do have secular trends and hence are nonstationary. Fortunately, most of these series can be changed to stationary time series by the differencing transformation.

If a series requires differencing, the first observation is subtracted from the second, the second observation from the third, and so on. The parameter d, the order of differencing, denotes the number of times we must perform the differencing operation in order to remove all secular trends from the data. Although differencing may transform a series so that it is suitable for ITSA, it does not remove the treatment effects that are to be assessed. (See McCain & McCleary, 1979, p. 236-238, for an elementary discussion of differencing.)

The estimation of the ARIMA (p,d,q) model may appear forbiddingly complex. Fortunately, there appear to be upper limits on the degree of complexity likely to be encountered in practical time-series analyses. For instance, although an ARIMA process could have more than two autoregressive components, higher order autoregressive models are rare. In an empirical investigation of approximately 100 time series involving social and behavioral sciences data by Glass et al. (1975), only 2% of the time series had more than one autoregressive term. Similarly, although higher order moving average models are possible, they also are rare (Glass et al., 1975). Furthermore, McCain and McCleary (1979) report they have never encountered a mixed process in an actual ITSA. Finally, 51% of the series investigated by Glass et al. (1975) required no differencing, and only 6% required differencing beyond the first order.

Q: How is the appropriate time-series model identified?

A: Model identification refers to determining the order of the model, that is, to determining the value of p, d, and q. Identification is based upon an examination of the autocorrelations and the *partial autocorrelations* calculated on the time-series data. Partial autocorrelations calculated on time-series data are analogous to partial correlations on typical temporally unordered data. For example, the lag-4 partial autocorrelation indexes the degree of predictability from an observation four intervals in advance with the interviewing observations held constant. (See Gottman & Glass, 1978, p. 213, for a further elaboration of partial autocorrelations.) The autocorrelations and the partial

autocorrelations are typically calculated for the first $N/4$ lags in the data set. The $N/4$ rule of thumb is used because higher order autocorrelations and partial autocorrelations become increasingly unstable, as they are based on progressively fewer observations. (For example, with $N = 75$ observations, the lag-50 autocorrelation would include only 25 pairs of observations. Observations beyond Z_{25} would have no observation separated by 50 lags with which to be paired.) These coefficients form the *autocorrelation function* (*ACF*) and the *partial autocorrelation function* (*PACF*). The ACF and the PACF are often displayed in graphs called *correlograms.*

The appropriate values of the ARIMA (p,d,q) model are inferred from the forms of the correlograms. The researcher should note, however, that the parameters of the ARIMA (p,d,q) model may change as a consequence of the experimental interaction (Stoline, Huitema, & Mitchell, 1980). If so, it may not be appropriate to infer values of p, d, and q from the correlograms computed from the entire time-series data set. Instead, Glass et al. (1975) suggested that the autocorrelation function and the partial autocorrelation function should be computed separately on the pre-intervention and the post-intervention data. Model identification is then based on the ACF and the PACF that are obtained by taking a weighted average of the pre- and post-intervention ACFs and the PACFs. McSweeney (1977) suggests that separate models should be identified for the pre-intervention data, the post-intervention data, and for the series as a whole. If inconsistent results are found, the model providing the most conservative estimate of the treatment effect should be chosen. McSweeney's method may prove troublesome in two respects (See Stoline, Huitema, & Mitchell, 1980 for another possible alternative). First, the selection of the most conservative results from among three analyses would likely reduce the power of ITSA. Second, the presence of a treatment effect may produce the appearance of nonstationarity in the ACF and PACF functions calculated on the entire series. This, in turn, may induce the investigator to incorrectly difference the series or otherwise perform an improper analysis.

The first step in analyzing a correlogram is to establish the stationarity of the process. For stationary time-series processes, we expect the influence of past observations to decrease rapidly as the lag increases. Secular trends, on the other hand, tend to elevate the level of the ACF at all lags. Consequently, if the lag-1 autocorrelation is near ± 1.0 and the succeeding lag correlations do not die out within four or five lags, nonstationarity should be suspected. The observations should be differenced and the ACF and PACF recomputed until a stationary pattern is achieved.

The reader should be warned, however, that it can be difficult to correctly difference a time series with small N in the presence of strong serial dependency. Extreme care should be exercised in these circumstances because, as Padia (Note 2) points out, errors in differencing can be critical in ITSA.

When the data have been satisfactorily differenced, the resulting ACF and PACF correlograms are examined for evidence of the order of the MA and AR components. The first hypothesis to be explored is that the process is white noise. Because white noise is by definition uncorrelated, white noise ACFs and PACFs should be zero for all lags. The ACF correlogram for the white noise data in Figure 1, Panel A, is shown in Figure 2, Panel A. The autocorrelations fluctuate around zero, as expected. The same is true for the white noise PACF shown in Figure 3, Panel A. It is possible, however, that sampling finite realizations of white noise may produce some correlograms containing chance significant "spikes." Consequently, we protect ourselves against misidentification by testing the ACF as a whole for evidence of significant serial dependency. This test is often made with the Box-Piece test, sometimes called the Q-statistic (Box & Jenkins, 1970; McCain & McCleary, 1979). The value

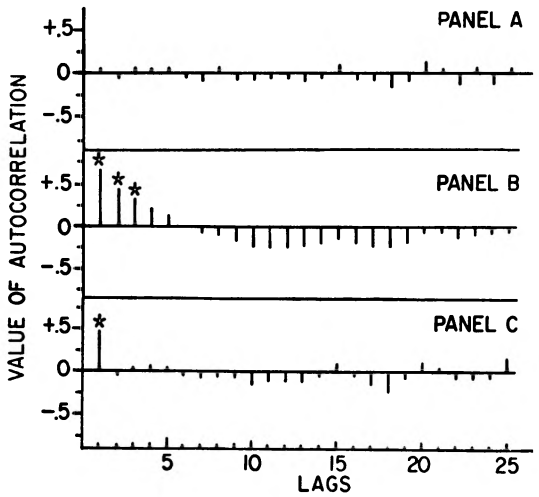

Fig. 2. Correlograms for the autocorrelation function (ACF) of (A) the white noise function, (B) the autoregressive process, and (C) the moving average process of Figure 1. Asterisk indicates that the value of the autocorrelation exceeds two standard deviations.

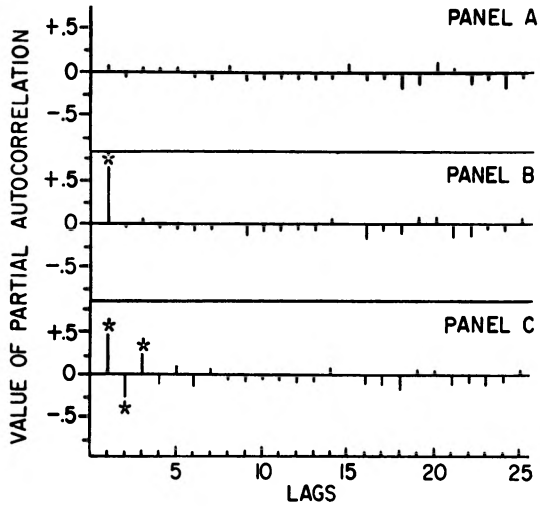

Fig. 3. Correlograms for the partial autocorrelation function (PACF) for (A) white noise process, (B) autoregressive process, and (C) moving average process of Figure 1. Asterisk indicates that the partial autocorrelation exceeds two standard deviations.

of Q for the ACF shown in Figure 2, Panel A, for the white noise process is clearly nonsignificant ($Q(25) = 13.7$, $p > .95$).

If there is evidence of significant autocorrelation, we proceed to identify the orders of the AR and MA components. We begin with the simplest possible models, ARIMA (1,0,0) or ARIMA (0,0,1), and do not proceed to more complex alternatives until such models have been proven inadequate. Detailed discussion of complex ARIMA (p,d,q) model identification including mixed models in which both p and q take on nonzero values, and models involving differencing (e.g., where d takes on a nonzero value) can be found in the standard reference sources described in the final section of this paper. We will discuss only the identification of AR(1) and MA(1) models for the sake of clarity. For order-1 processes, the ACF and PACF correlograms appear in two complementary forms. The correlation functions either show a large significant spike at lag 1, and then cut off to zero; or they show a large value at lag 1 and then decay slowly in the succeeding lags.

Individual autocorrelations and partial autocorrelations can be tested for significance by taking the ratio of these statistics to their standard deviations. The standard deviation of the lagged-k autocorrelation is equal to

$$[1/N(1 + 2\sum_{j=1}^{k} r_j^2)]^{1/2} \text{ where } \sum_{j=1}^{k} r_j^2$$

is the sum of the squared values of the first through the kth autocorrelations. The standard deviation for all partial autocorrelations is taken as $(1/N)^{1/2}$.

For the AR(1) process, the ACF (Panel B, Figure 2) shows the decay pattern and the PACF (Panel B, Figure 3) shows the spike. The evidence for the MA(1) function is the mirror image of the AR(1): the ACF (Panel C, Figure 2) reveals a spike at lag 1, while the PACF (Panel C, Figure 3) shows the decay pattern. The sign of the ϕ or the θ coefficients within these models may be inferred from the directions of spike values (— or +) and the presence or absence of an alternation of sign in the decaying ACF or PACF. The positive decay in the ACF and positive spike in the PACF of

the AR(1) correlograms (Panel B, Figure 2; Panel B, Figure 3) indicate a positive ϕ value. The positive spike in the ACF and alternating decay in the PACF of the MA(1) correlograms (Panel C, Figure 2; Panel C, Figure 3) indicate a positive θ value. When an ARIMA (p,d,q) model has been tentatively identified, the values of the ϕ and/or θ coefficients in the model are estimated on a computer. The model is then tested for adequacy.

Q: How does one know if the appropriate ARIMA model has been identified?
A: There are a number of ways of determining whether the appropriate ARIMA model has been identified. If errors have been made in model identification, they may be discovered by examining the printouts of the weights of the autoregressive and moving average terms. These weights must be within certain bounds. These bounds, called the stationarity and invertibility bounds (see Glass et al., 1975), are analogous to the bounds of the ordinary product moment correlation; that is, the value of r cannot exceed $+1.0$ or be less than -1.0. Just as an obtained r outside these limits indicates the presence of an error, so too do values of ϕ and θ that lie outside their bounds indicate an error in model identification. Model identification errors are also disclosed when the value of ϕ or θ required by the model are not significantly different from zero. For example, if an ARIMA (1,0,1) model has been identified, then the autoregressive term ϕ and the moving average term θ must be significantly different from zero. If they are not, then the model has been misidentified, and a revised model must be formulated.

The final step in troubleshooting or diagnosing model identification errors occurs when the ACF and the PACF calculated on the *residual* scores are examined. These residual scores are the portion of the original scores remaining after the estimated autoregressive and moving average components have been removed. These residual scores should now be serially independent, and should resemble the realization of a white noise process. That is, the ACF and the PACF calculated on the residual scores should have no spikes at early lags and should *not* be significantly different from zero as tested by, for example, the Q-statistic. If the ACF and the PACF differ from zero, then the whole procedure of model identification, estimation, and diagnosis must be repeated until an acceptable model has been identified and the weights (i.e., ϕ and/or θ) for the parameters of the model have been estimated. When this procedure has been completed, the residual scores from which serial dependency has been removed can then be tested for the presence of treatment effects.

Q: How are treatment effects tested?
A: After an appropriate model has been fitted to the stochastic component of the time-series data, intervention effects can be tested. The intervention components in ITSA are sometimes referred to as "transfer functions." Some authors, e.g., Anderson (1976) and Box and Jenkins (1970), use *transfer function* to describe what we have called the *stochastic component* of the time series model. Our usage follows McCain and McCleary (1979). These functions transfer the level, the slope, or both the level and slope of the series from one state during a time period (e.g., the baseline phase) to another state during a subsequent time period (e.g., the treatment phase). See, for example, Jones et al. (1977), Kazdin (1976), McCain and McCleary (1979), and Glass et al. (1975).

On the simplest level, one could model an abrupt change in level as the time series moves from the baseline phase to the experimental phase. Panel A of Figure 4 illustrates how such a transfer function might appear. It is often the case, however, that an intervention affects a subject gradually. One of the advantages of ITSA is that an asymptotic rise to a new level as a result of treatment can be explicitly modeled and tested, as illustrated in Panel B of Figure 4. Conversely, an investigator could model an intervention that results in an instantaneous improvement that is not sustained.

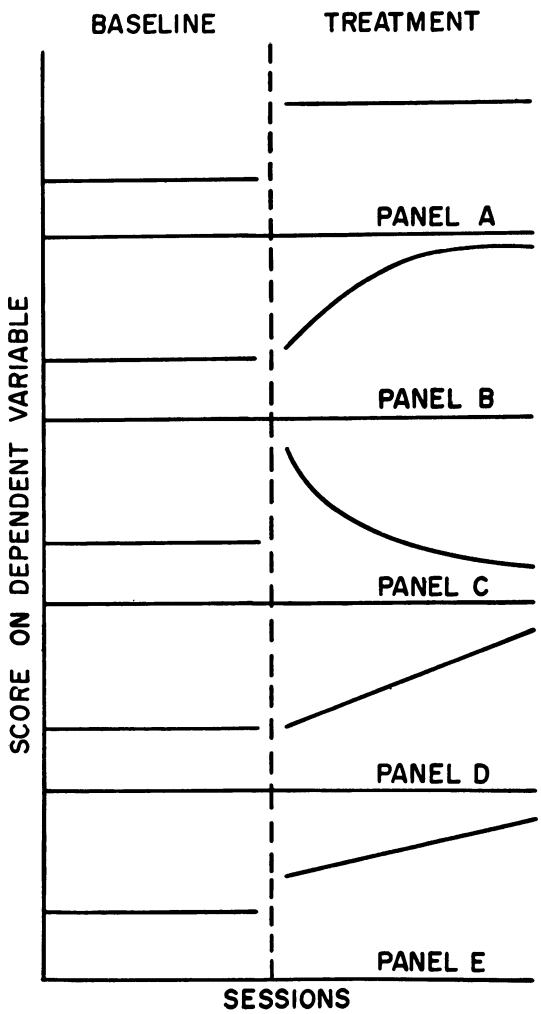

Fig. 4. Plots of transfer functions that can be modeled in ITSA. In each case the beginning of change coincides with the initiation of treatment. Panel A depicts an abrupt change in level. Panel B shows an asymptotic rise. Panel C illustrates an abrupt change in level which subsequently dies away. Panel D shows a change in slope. Panel E illustrates a change in both level and slope.

Panel C illustrates the transfer function that induces an abrupt change in the level of the time-series process which subsequently dies away. Or, the effect could be a change in the slope of a function with a constant level, as illustrated in Panel D. Finally, ITSA allows the researcher to model the effect of an intervention with a combination of transfer functions. Panel E illustrates a combination of an abrupt change in level with a change in slope. The statistical tests on transfer functions follow ordinary statistical procedures: Model parameters are estimated and these estimates are tested (e.g., by t tests) to determine if the obtained values fall within or outside the boundary that defines statistical significance. For details on the procedures of transfer function modeling and testing, see McCain and McCleary (1979) or Glass et al. (1975).

It should be noted that the statistical test of change from pre- to post-intervention is not by itself a test of the causal relationship between a treatment manipulation and a dependent variable. The extent to which "cause" can be inferred is dependent upon design and measurement factors as well as the results of statistical tests.

In summary, the modeling-testing strategy described here includes four steps: (a) tentative identification of an ARIMA (p,d,q) model from the ACF and PACF; (b) estimation of the values of the autoregressive and moving average parameters (ϕ and θ) for the tentatively selected stochastic model; (c) diagnostic assessment of the adequacy of the stochastic model selected in the first step (note: if the model selected is improbable or otherwise inappropriate, the identification/estimation/diagnostic procedure continues until an acceptable ARIMA model is found); and finally, (d) modeling a transfer function to describe the treatment effects. The transfer function can be tested using conventional statistical procedures.

THE INTERPRETATION AND PRESENTATION OF RESULTS FROM ITSA

Q: What conclusions should be drawn if the results of a visual analysis and the results of an ITSA are inconsistent?

A: Jones et al. (1978) have presented evidence that visual analysis and ITSA may provide inconsistent results, particularly when treat-

ment effects are not dramatic and when serial dependency is substantial (i.e., when $r_1 > .75$). If these conditions are frequent in applied behavioral studies, then disagreement between these two decision aids may be common.

In discussing the relative merits of visual analysis and ITSA, one must be careful not to present ITSA as superior merely because it is objective. Both visual and statistical analyses require subjective judgments by the investigator. Visual analysis is subjective because the patterns it finds in the data are neither measured nor compared against any objective criterion. But, as we have indicated, ITSA also requires the analyst to make judgments about the appropriate stochastic model for the time-series data on the basis of patterns "seen" in the correlograms. This is a procedure that has not been automated and certainly requires a subjective contribution from the researcher. This may be a disguised virtue. The statistical facilities available in "canned" programs on computers may lead us to forget that all statistical methods have underlying models that must be implicitly or explicitly fitted to the data in the course of analysis. ITSA requires strong assumptions that must be explicitly stated. Moreover, ITSA provides a diagnostic procedure for determining the confidence of such judgments. Visual analysis differs, in that because some of the constituent parts of the judgment based upon visual inspection cannot be stated, there is simply no way to determine the certainty with which such judgments are made.

There are clearly two ways in which visual and statistical judgmental aids may disagree. First, an ITSA may produce evidence of an experimental effect that is not supported by inspection of the graphed time-series data. Because, in this case, it seems unlikely that an effect of intervention would be judged to have immediate clinical significance, many applied researchers would be inclined to discount the results of the ITSA. However, such a result may suggest the presence of a treatment worthy of additional investigation either through replication or by means of an improved experimental design.

The second possible disagreement, where treatment effects seem apparent but ITSA indicates that the null hypothesis cannot be rejected, may place the analyst in a more difficult quandary. Here, however, consideration of the probable relative validity of visual inspection and ITSA may indicate a choice between the two conflicting judgments.

We know something about the conditions under which ITSA will produce unstable results and something about the conditions that will similarily erode the reliability of visually based judgments. However, because investigations of the trustworthiness of these two judgmental aids are few in number, we can only give tentative guidelines concerning how strongly to weigh the evidence of a time-series analysis that disagrees with an investigator's visual judgment.

A frequent difficulty in the use of ITSA is brevity of the time series. Short series have two undesirable consequences: model-fitting (particularly differencing) is performed with less confidence in the adequacy of the model, and statistical tests of intervention effects are less likely to detect real changes. Obtaining the correct level of differencing is particularly important because inappropriate differencing can lead to erroneous statistical test results (Padia, Note 2). Under-differencing leaves serial dependency in the series and over-differencing introduces unwanted serial dependency into the series. Padia recommends that the number of observations in the baseline be not less than 50 and that 50 be the minimum number of postintervention points as well. Padia (Note 2) states, "For the smaller number of points it is difficult to determine the degree of differencing required to obtain stationarity, since a 'wandering' over the short run may be either highly correlated stochastic fluctuations in the stationary series or the 'drifting' of a non-stationary series" (p. 144).

Typically, other recommendations have been 50-100 observations within a single phase (Box

& Jenkins, 1970; Gottman & Glass, 1978). Although more observations are better than fewer observations, the question, "How many data points are necessary to perform an ITSA?" is not answerable in its most general form. Under some circumstances, substantially fewer than 50 observations may be appropriate. For example, if we can assume a simple model such as an ARIMA (1,0,0), and if this model is a reasonable fit, the Glass et al. (1975) algorithm is quite powerful with as few as 20 baseline and 20 postintervention observations. It should be pointed out that nothing is known about the reliability of visual analysis in these circumstances. It is a plausible though untested hypothesis that visual analysis is also less reliable with small numbers of observations. Nevertheless, with time series that include few observations, particularly those in which visual analysis indicates a treatment effect that is not supported by the results of ITSA, greater credence may be given to the results of the visual analysis. If, on the other hand, a time series contains many observations, and the data are highly serially correlated, variable, and have baseline trends—all conditions that seem to present problems for visual decision makers—the results of the interrupted time-series analysis would be afforded greater confidence than would the results of the visual analysis. Whatever the circumstances, when faced with uncertain results, the investigator would likely replicate the study before sharing the results with his or her peers.

Q: Will the necessity or desirability for numerous observations preclude the use of ITSA by applied behavior analysts?

A: Decisions about implementing and withdrawing treatment intervention are made on the basis of the pattern of the data. These decisions determine the length of phases, and hence the number of observations available. Fortunately, overlapping considerations lead to decisions to extend phases and also suggest that ITSA would be a useful decision aid. These considerations include small behavioral changes, unstable baselines, and variable data. Thus, more abundant data may frequently be available when they are most needed.

Aside from this general consideration, specific characteristics of commonly used intrasubject-replication designs may limit the applicability of ITSA. For example, multiple-baseline designs characteristically include components with brief baseline or treatment phases, and changing criterion designs may often include brief subphases. In these cases, sufficient observations may not be available to conduct an entirely adequate ITSA, even if the pattern of the data suggests that a judgmental aid other than visual analysis would be useful. Even here, however, a second, albeit flawed, decision aid may be preferable to a single flawed decision aid.

In other cases, ITSA may be of limited value because conclusions are drawn in ways that depart from straightforward comparisons of changes from one phase to another. For example, time-series analysis is of unknown relevance in assessing the comparative effectiveness of two or more *concurrent* treatments as these might be examined in a simultaneous treatment design or a multiple-element design. These apparent incompatibilities between ITSA and individual subject designs are not meant to discount the usefulness of interrupted time-series analysis, but to place in perspective its likely role.

Q: What technical information should be included in a manuscript using ITSA?

A: Aside from the raw time-series data, the technical information that should be included in a manuscript using ITSA is of two kinds: (a) the information used in determining the parameters of the time-series model applied to the stochastic components of the data, and (b) the summary of the procedures used to test treatment effects. The information used in model fitting should include the ARIMA (p,d,q) model fitted to the stochastic component of the time-series data. The correlograms of the ACF and

Table 1
Selected Applied and Technical References on ITSA

References	*Comments*
Applied	
Deutsch and Alt (1977)	Controversial application of ITSA to assess the effects of gun-control legislation on gun-related crimes in Boston. See Hay and McCleary's (1979) critical evaluation and Deutsch's (1979) vigorous reply.
Gottman and McFall (1972)	Application of ITSA to evaluate the effects of self-monitoring on the school-related behavior of high school dropouts.
McSweeny (1978)	Description of the effects of a response-cost procedure on the telephoning behavior of people in Cincinnati. The statistical analysis performed on the data are described in McCain and McCleary (1979).
Schnelle, Kirchner, McNees, and Lawler (1975)	Application of ITSA to assess the effectiveness of saturation patrols on burglary rates.
Technical	
Jones et al. (1977)	A nontechnical presentation of ITSA.
Kazdin (1976)	A discussion of the problems of serial dependency and several technical alternatives, including ITSA.
McCain and McCleary (1979)	The clearest technical introduction to ITSA.
Gottman and Glass (1978)	A restatement and updating of Glass et al. (1975).
Nelson (1973)	A summary of the Box-Jenkins theory in practice. Should be readable for those with a strong background in multiple regression. Applications illustrate the use of the ESP computer package.
Hibbs (1974)	Contains a good discussion of the problems of serial dependency in statistical tests of time series data. Requires a familiarity with matrix algebra.
McCleary and Hay (1980)	A comprehensive applied treatment of the Box-Jenkins method designed for behavioral and social scientists.
Glass et al. (1975)	Summary of the Box-Tiao method and the discussion of its application to behavioral and evaluation research.
Anderson (1976)	A well-written, but mathematically sophisticated digest of Box-Jenkins.
Box and Tiao (1965)	A difficult but fundamental article on the analysis of interventions in time series.
Box and Jenkins (1976)	A treatise in mathematical statistics. The source for most of the other references.

Note. Technical articles are listed in order of difficulty. The present article is of approximately equal difficulty to the articles by Kazdin (1976) and by McCain and McCleary (1979).

PACF, and the test of whether the chosen model generated uncorrelated residuals could also be provided to allow the reader the opportunity to review the basis for the investigator's judgments about these model parameters. The details of model fitting could be elaborated further in manuscripts using ITSA, but much more detail than this may confuse readers. The set of technical information concerning the inferential tests performed on the data to assess treatment effects should include the t values for changes in level and slope (drift) with appropriate reference to degrees of freedom and significance levels.

SOURCES OF INFORMATION ON ITSA

Q: Where can I find out more about ITSA?
A: As with other complex statistical techniques, the use of ITSA requires that the analyst get a working knowledge of its underlying mathematics. Increasing numbers of articles, chapters, and books are being published that use interrupted time-series analysis to answer a substantive behavioral question, to describe how a time-series analysis can be used, or to discuss technical issues related to the use of ITSA. This material varies substantially in difficulty, and

unless one enters this literature at an appropriate level of difficulty, the experience can be punishing. To promote successful avoidance of punishing experiences, we have prepared an annotated set of references (Table 1) with the technical references graded in approximate level of difficulty.

Q: What are some computer programs useful for performing ITSA?
A: Each of the four steps in ITSA is performed with the aid of appropriate computer programs. To assist the reader in locating computer software for ITSA, we present a table of existing and forthcoming programs. Automatic Forecasting System and TMS are available from their authors, and IMSL is widely supported at university computer installations. BMDP and SPSS are also widely available, but their time-series subprograms are forthcoming or only recently distributed.

Two types of programs are listed in Table 2. Automatic Forecasting System, TMS, BMDP, and SPSS are "package" programs, designed to perform the entire range of ITSA functions. They do not require any programming by the user. IMSL is not a package, but rather a library of FORTRAN subroutines. IMSL provides several subroutines useful for modeling the stochastic process within the time-series data, but they must be called by a program written by the user. IMSL does not provide subroutines explicitly designed for modeling the effect of an intervention in ITSA.

SUMMARY AND CONCLUSION

Since the recognition that serial dependency (Jones et al., 1978) and other individual subject data characteristics (e.g., DeProspero, & Cohen, 1979) lead to unreliable, and hence invalid, visually based assessments of behavioral change, the need for alternative or supplementary decision aids has become clear. In this paper we have attempted to familiarize applied behavioral researchers with one such technique, interrupted time-series analysis.

Although ITSA has certain clear advantages, we do not suggest that this technique should be applied uniformly to all time-series data. Some patterns of results are clearly detectable by visual inspection and do not require supplemental decision aids, e.g., a lengthy baseline phase in which the data are stable and have zero slope followed by a lengthy intervention phase in which the data are stable and show an abrupt

Table 2

Computer Programs for ITSA

Name	*Authors*	*Language*	*Comments*
Automatic Forecasting Systems	Pack (1978)	FORTRAN	Originally developed under the supervision of Box and Jenkins. For sophisticated users.
TMS	Bower, Padia, and Glass (1974)	FORTRAN	Developed at the University of Colorado in conjunction with Glass et al., (1975). Requires some sophistication.
IMSL	IMSL (1979)	FORTRAN	"International Mathematical Statistics Library." Sophisticated subroutines for the experienced FORTRAN programmer. Relevant subroutines include FTRDIF, FTAUTO, FTCMP, FTMPS, and FTMXL.
BMDP	——	——	Forthcoming as BMDP2T. From an inspection of a preliminary version of the documentation, this appears to be a powerful, but user-oriented and accessible package.
SPSS	——	——	Forthcoming.

change in level with zero slope. Moreover, ITSA may be unwarranted because either the cost of making an incorrect visually based judgment exceeds the cost of performing the analysis, or because the experimental design used is unsuitable for ITSA. Aside from these qualifications, we recommend that applied behavior analysts strongly consider supplementing visual inspection of their data displays with ITSA. By conducting interrupted time-series analyses, they will (a) gain information about the properties of their time-series data that might not otherwise be available, (b) improve the quality of their future visually based judgments, (c) assess more adequately certain types of treatment effects such as changes in slope, and (d) protect themselves against both the false acceptance of nonexistent effects and the false rejection of existent treatment effects. These advantages should compensate researchers for the effort required to perform the analysis.

REFERENCE NOTES

1. Kennedy, R. E. *The feasibility of time series analysis of single case experiments.* Unpublished manuscript, The Pennsylvania State University, 1976.
2. Padia, W. L. *The consequence of model misidentification in the interrupted time-series experiment.* Unpublished doctoral dissertation, University of Colorado, 1975.

REFERENCES

Anderson, O. D. *Time series analysis and forecasting: The Box-Jenkins approach.* London: Butterworth, 1976.

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

Bower, C. P., Padia, W. L., & Glass, G. V. *TMS: Two FORTRAN IV programs for the analysis of time-series experiments.* Boulder: Laboratory of Educational Research, University of Colorado, 1974.

Box, G. E. P., & Jenkins, G. M. *Time-series analysis: Forecasting and control.* San Francisco: Holden-Day, 1970.

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

Box, G. E. P., & Tiao, George C. A change in level of a non-stationary time series. *Biometrika,* 1965, **52,** 181-192.

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

Deutsch, S. J. Lies, damn lies, and statistics: A rejoinder to the comment by Hay and McCleary. *Evaluation Quarterly,* 1979, **3,** 315-328.

Deutsch, S. J., & Alt, F. B. The effect of Massachusetts' gun control law on gun-related crimes in the city of Boston. *Evaluation Quarterly,* 1977, **1,** 543-568.

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.

Gottman, J. M., & McFall, R. M. Self-monitoring effects in a program for potential high school dropouts: A time-series analysis. *Journal of Consulting and Clinical Psychology,* 1972, **39,** 273-281.

Hay, R. A., & McCleary, R. Box-Tiao time series models for impact assessment: A comment on the recent work of Deutsch and Alt. *Evaluation Quarterly,* 1979, **3,** 277-314.

Hibbs, D. A., Jr. Problems of statistical estimation and causal inference in time-series regression models. In H. L. Costner (Ed.), *Sociological methodology 1973-1974.* San Francisco: Jossey-Bass, 1974.

IMSL, *IMSL* (Vol. 2; 7th ed.). Houston: IMSL, 1979.

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. Effects 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.* Oxford: Pergamon Press, 1976.

McCain, L. J., & McCleary, R. The statistical analysis of the simple interrupted time-series quasi-experiment. In T. D. Cook & D. T. Campbell, *Quasi-experimentation: Design & analysis issues for field settings.* Chicago: Rand McNally, 1979.

McCleary, R., & Hay, R. A., Jr. *Applied time series analysis for the social sciences.* Beverly Hills, CA: Sage Publications, 1980.

McSweeney, A. J. Time series analysis and research in behavior modification: Some answers: *AABT Newsletter,* 1977, **4,** 22-23.

McSweeney, A. J. Effects of response cost on the behavior of a million persons: charging for directory assistance in Cincinnati. *Journal of Applied Behavior Analysis,* 1978, **11,** 47-51.

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

Nelson, C. R. *Applied time series analysis.* San Francisco: Holden-Day, 1973.

Pack, D. J. *A computer program for the analysis of time series using the Box-Jenkins philosophy.* Hatboro, Pa.: Automatic Forecasting Systems, 1978.

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.

Scheffé, H. *The analysis of variance.* New York: Wiley, 1959.

Schnelle, J. F., Kirchner, R. E., McNees, M. P., & Lawler, J. M. Social evaluation research: The evaluation of two police patrolling strategies. *Journal of Applied Behavioral Analysis,* 1975, **8,** 353-365.

Sidman, M. *Tactics of scientific research.* New York: Basic Books, 1960.

Stoline, M. R., Huitema, B. E., & Mitchell, B. T. Intervention time-series model with different pre- and postintervention first-order autoregressive parameters. *Psychological Bulletin,* 1980, **88,** 46-53.

Tukey, J. W. *Exploratory data analysis.* Reading, Mass.: Addison-Wesley, 1977.

Received October 24, 1979
Final acceptance March 27, 1980