Psychological Bulletin
1969, Vol. 72, No. 4, 299–306

DESIGN AND ANALYSIS OF RESEARCH USING TIME SERIES

JOHN M. GOTTMAN, RICHARD M. McFALL,[1] AND JEAN T. BARNETT

University of Wisconsin

A time-series methodology is developed for approaching data in a range of research settings. A design package is presented using the time series as a method to eliminate major sources of rival hypotheses. A mathematical model is offered which maximizes the utility of time-series data for generating and testing hypotheses. Special considerations in the application of the model are discussed.

The purpose of the present paper is to present a methodological approach to such research areas as psychotherapy, education, psychophysiology, operant research, etc., where the data consist of dependent observations over time. Existing methodologies are frequently inappropriate to research in these areas; common field methodologies are unable to control irrelevant variables and eliminate rival hypotheses, while traditional parametric laboratory designs relying on control groups are often unsuitable.

New data-analysis techniques have made possible the development of a different methodological approach which can be applied in either the laboratory or in natural (field) settings. This approach is responsive to ecological considerations (Willems, 1965) while permitting satisfactory experimental control. Control is achieved by a network of complementary control strategies, not solely by control-group designs.

The Use of Time Series in Design

The most persuasive experimental evidence comes from a triangulation of research designs as well as from a triangulation of measurement processes. The following three designs, when used in conjunction, represent such a triangulation: (*a*) the one-group pretest-posttest design; (*b*) the time-series design; and (*c*) the multiple time-series design. These designs need not be applied simultaneously; rather, they form a complementary network of designs, each meeting different research demands by eliminating different sources of rival hypotheses. A detailed evaluation of each of these designs has been presented elsewhere (Campbell, 1967; Campbell & Stanley, 1963).

The One-Group Pretest-Posttest Design

This design, although inadequate when used alone, makes a significant and unique contribution to the total design package. It provides an external criterion measure of the outcome of a programmed intervention. Each subject serves as his own control, and the difference between his pre- and posttest scores represents a stringent measure of the degree to which "real life" program goals have been achieved.[2] For example, the ultimate success of psychotherapy is best evaluated in terms of extratherapeutic behavior change. This design, then, documents the fact of outcome-change without pinpointing the process producing the change.

The Time-Series Design

This design involves successive observations throughout a programmed intervention and assesses the characteristics of the change process. It is truly the mainstay of the proposed design package because it serves several simultaneous functions. First, it is *descriptive.* The descriptive function of the time series is particularly important when the intervention extends over a considerable time period. The time series is the only design to furnish a continuous record of fluctuations in the experimental variables over the entire course of the program. Such record keeping should constitute an integral part of the experimental pro-

[1] Requests for reprints should be addressed to Richard McFall, Department of Psychology, Charter at Johnson, University of Wisconsin, Madison, Wisconsin 53706.

[2] Whereas Campbell (1967) asserts that *experimental mortality* is controlled by this design, mortality may, in fact, act as a source of variance. A differential response to treatment may systematically influence who drops out of the experiment.

gram; problems of reactivity (Webb, Campbell, Schwartz, & Sechrest, 1966) area voided by incorporating the measurement operations as a natural part of the environment to which one wishes to generalize.

Second, the time-series design functions as an *heuristic* device. When coupled with a carefully kept historical log of potentially relevant nonexperimental events, the time series is an invaluable source of post hoc hypotheses regarding observed, but unplanned, changes in program variables.[3] Moreover, where treatment programs require practical administrative decisions, the time series serves as a source of hypotheses regarding the most promising decisions, and later as a feedback source regarding the consequences and effectiveness of such decisions.[4]

Finally, the time series can function as a quasi-experimental design for planned interventions imbedded in the total program when a control group is implausible. Figure 1 depicts a time-series experiment with an extended intervention; of course, in some cases, the intervention might simply be a discrete event.

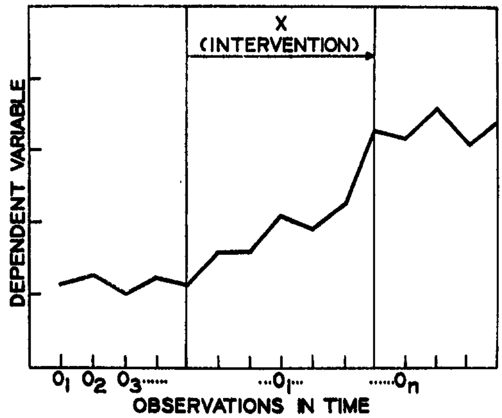

FIG. 1. A time series with an extended intervention.

A time-series analysis must demonstrate that the perturbations of a system are not uncontrolled variations, that is, noise in the system. It is precisely this problem of partitioning noise from "effect" that has discouraged the use of time series in the social sciences. Whereas the uncontrolled variations in *physical* science experiments can be assumed to be small when compared to experimental effects, the uncontrolled variations encountered in *social* science experiments often surpass the experimental effects.

Ezekiel and Fox (1966) in a discussion of the history of time series in the social sciences say that:

> In the early and middle 1920's many researchers were completely unaware of problems connected with the sampling significance of time-series. Then, under the (partly misinterpreted) influence of such articles as Yule's (1926) on "nonsense correlations," it became fashionable to maintain that error formulas simply did not apply to time-series. There was some implication that reputable statisticians should leave time-series alone. . . . During the 1930's, therefore, some research workers continued to apply regression methods to time-series, but with considerable trepidation [p. 325].

The present paper contends that such a reluctance to use the time-series design on statistical grounds is no longer necessary. In a subsequent section of this paper, appropriate statistical procedures will be presented rendering the time series useful once again.

In summary, this design is more capable of eliminating plausible rival hypotheses for data than was the one-group pretest-posttest design[5]; it serves as a technique for generating an overall description of programmatic change, and it functions as a source of hypotheses regarding the nature of the *process* of change.

The Multiple Time-Series Design

This design is basically a refinement of the simple time series. It is yet a more precise method for investigating specific program hypotheses because it allows the time series of the experimental group to be compared with that of a control group. As a result, it offers a greater measure of control over unwanted sources of rival hypotheses.

The multiple time series is the first component of the design package to require a comparison group. Use of a comparison group raises the

[3] Such a log could provide critical-incident data (Flanagan, 1954).

[4] A data-overload situation, an ever present possibility, should be avoided by limiting observation to only a select set of salient variables.

[5] As with the one-group pretest-posttest design, and for the same reasons, the present authors disagree with Campbell's (1967) assertion that this design controls for mortality. It is controlled, however, by the time-lagged multiple time-series design.

practical question of group-selection procedures and in some situations, such as psychotherapy, it also raises the ethical problem of using no-treatment or minimal treatment groups. The time series suggests two approaches to these problems. First, one may use a statistically nonequivalent comparison group in which subjects have not been randomly assigned to treatment and comparison groups. Usual problems with such a procedure can be solved with special techniques of data handling. Second, and preferably, one may use a time-lagged control group where the intervention is temporarily withheld from one group of subjects but not another (see Figure 7, p. 305). This procedure also provides information on whether the effect of an intervention is tied to a specific time.

The Analysis of Time-Series Data

The data resulting from the best of experimental designs is of little value unless subsequent statistical analyses permit the investigator to test the extent to which obtained differences exceed chance fluctuations. The above design package, with its emphasis on time-series designs, is a realistic possibility only because of recent developments in the field of mathematics. Appropriate analysis techniques have evolved from work in such diverse areas as economics, meteorology, industrial quality control, and psychology. Historically, the time-series design has been neglected due to the lack of such appropriate analytical techniques. Two statistical methods for solving the problem of time-series analyses are presented below.

Curve Fitting

Curve fitting is the simplest and best known approach to the analysis of time-series data. It involves fitting the data to the *least squares* straight lines. The data are divided into two classes, the class of observations, or points, which precede the intervention and the class of those which follow the intervention. One straight line is used to fit the first class of points and another to fit the second class. The difference in slope and intercept of both lines projected to X (the point of intervention) is then calculated and an appropriate test of significance is performed. One such significance test is given by Mood (1950). Figure 2 illustrates the curve-fitting procedure.

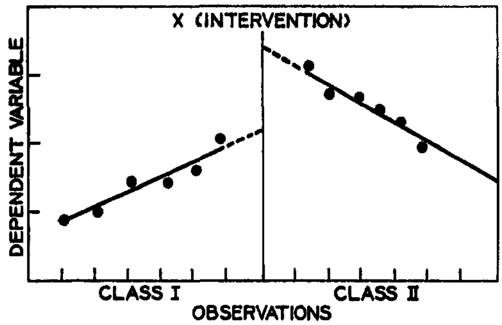

Fig. 2. Linear curve fitting.

There are at least two problems in using curve fitting with time series. First, the assumption of linearity is often inappropriate. When one attempts to fit a straight line to a set of observations in which the "underlying" relationship is not linear, one may find that the residuals are not randomly distributed; in effect, the straight line accounts for only a fraction of the total variance. In an attempt to overcome this problem, Alexander (1946) provided a method for calculating the trend away from linearity. If the trend is found to be significant, one can specify the *nature* of the nonlinear trend by using Grant's (1956) procedure for calculating the higher order trends (i.e., the quadratic, cubic, quartic, etc., components of the nonlinear trend). One may then calculate the contributions of these higher order terms to the total variance. When successive contributions become insignificant, then one can truncate the fitting procedure and fit the data to a set of orthogonal polynomials by a least squares procedure (Grant, 1956). However, this solution is often unsatisfactory because as higher order trends are calculated, an increasing number of degrees of freedom are sacrificed.

The second weakness in the curve-fitting approach is its underlying assumption that the repeated observations are *independent* samples of a random variable. This assumption may be violated by the time-series design because repeated observations through time are often sequentially dependent (Holtzman, 1967). To justify the use of curve-fitting procedures one must argue that a sequentially dependent

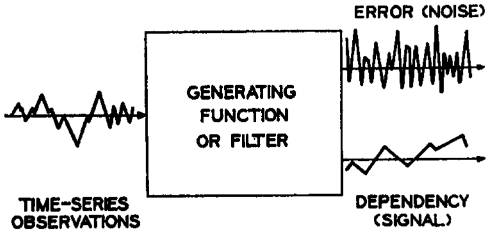

FIG. 3. Generating-function operation.

set of observations gives less information than a completely independent set. Using this argument, one can apply the Bartlett (1935) correction on the number of degrees of freedom.

Generating Function

The generating-function procedure, although less well known in the social sciences, is far more powerful than curve fitting for analyzing nonlinear, dependent time-series data, because it makes *positive* use of the dependency observations; the generating model is specifically derived from an analysis of such dependency. The generating-function procedure provides a solution to the problem of partitioning noise from effect. It not only clarifies the manner in which the time series is generated but also suggests how the time series might change as a function of different inputs. The dependent time series can be understood as consisting of a *signal* (the underlying dependency of the observations over time) which has been combined with *white noise* (error). The generating function, then, operates to separate the signal from the noise, as shown in Figure 3.

In the estimation problem,[6] the time series is given and a generating function must be found which breaks the series into two components—independent random fluctuations (nonsystematic error) and the remaining dependent, systematic variations. This problem is equivalent to investigating the nature of the time-series' dependency.

Stated mathematically, the problem is to estimate a function $F(D)$, such that the time series $x_t = F(D)\, e_t$, where e_t is error, and $F(D)$ is the function of a "shift" operator D, where $Dx_t = x_{t-1}$. To identify the operator one can investigate the nature of the time-series correlation structure. The correlation structure essentially tells us how well the series "remembers" its past history, that is, how strongly x_t depends on x_{t-1}, x_{t-2}, etc. To study the correlation structure of a time series, one calculates the *autocorrelation function*. This function is the correlation of a time series with itself, obtained by pairing observations t units apart ($t = 1, 2, \cdots$). This gives the serial correlation as a function of lag. A test for the significance of the autocorrelation function is given by Anderson (1942).

Two generating functions have found wide application in engineering, industrial, and economic time series. The first of these is the first-order moving-average function $x_t = e_t + a_1 e_{t-1} = (1 + a_1 D)\, e_t$, where a_1 is a constant. Here $F(D) = 1 + a_1 D$. In terms of the observations, by substitution, this equation can be shown to be equivalent to an "exponentially weighted moving average" of previous observations, plus an error term. This says that observation x_t remembers the previous observation most and the other observations a bit less. The closer the observation is to x_t, the more influential it is in predicting x_t.

A second commonly used generating function is the autoregressive process. The first-order autoregressive process is $x_t = b_1 x_{t-1} + e_t$; or $(1 - b_1 D)x_t = e_t$; or $x_t = \frac{1}{1 - b_1 D} \cdot e_t$.

This states that the next value of the time series is given by a constant b_1 times the previous value, plus an unpredictable noise e_t. Examples of such time series are given in Figure 4, with $b_1 = 0.9$ and -0.9, respectively.

Two models for the generating function $F(D)$ have been presented: the first-order moving-average model and the first-order autoregressive model. In general, a model is called a moving-average model if $F(D)$ is a polynomial in D, and an autoregressive model if $F(D)$ is the inverse of a polynomial in D. The basic problem of fitting a model to the data can be divided into three parts: (*a*) *Identification*—using the data or any other additional knowledge to suggest whether the series can be described as moving average, autoregressive, or perhaps a mixed model; (*b*) *Estimation*—using the data to estimate the parameters of $F(D)$; and (*c*)

[6] The estimation problem is one step removed from linear curve fitting because even a linear generating function can generate a nonlinear time series (Wold, 1965).

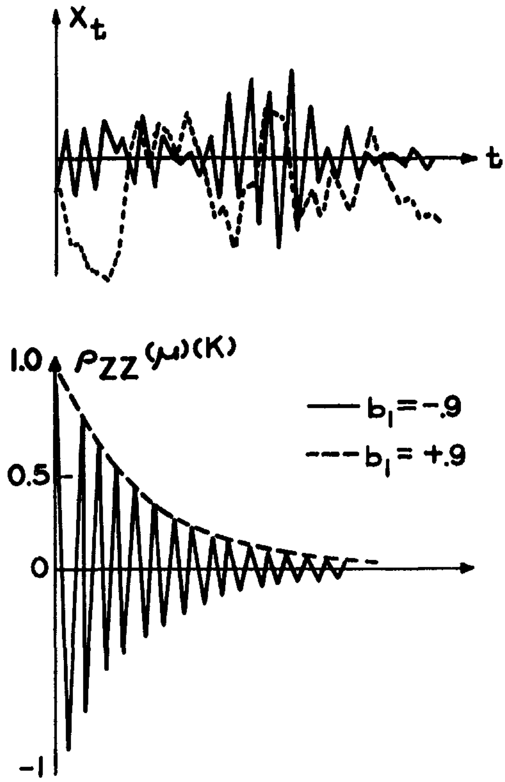

FIG. 4. A realization and the autocorrelation function of a discrete first-order autoregressive process (after Watts, 1967).

Diagnostic Checking—estimation of the residuals from the fitted model for lack of randomness and the use of this information to modify the model. An excellent discussion of this fitting procedure is given by Watts (1967) and Box, Jenkins, and Bacon (1967).

The Exponentially Weighted Moving-Average Model

Most time series in industrial, economic, or engineering applications use many observations (about 200 before statisticians feel comfortable), thus permitting refined determinations of the model and its parameters. However, in the social sciences there tend to be fewer observations, thus simpler models are warranted.

Experience has shown that the modified moving-average model is quite sufficient for most problems, even those using a large number of observations. As Coutie (1962) said, "The only justification for such a relatively simple procedure is that we have applied it to a wide range of [problems] and that it works well [pp. 345–346]." The moving average can be considered an approximate autoregressive process and vice versa (Watts, 1967). However, Box and Tiao (1965) said,

> The fact that . . . the weight function $F(D)$ . . . is uniform emphasizes the restrictiveness of the autoregressive model. Specifically, our results imply that this model is only acceptable if observations near the beginning and the end of the time-series have as much weight in the estimation [of a shift in the time series following an event E] as those close to the event E. In many industrial and economic applications, it seems much more reasonable to suppose that as we move away from E, the observations should become less and less informative about [the shift] [p. 188].

This is precisely what we find with the exponentially weighted moving-average model.

The exponentially weighted moving-average model (EWMA) is a simple dynamic model which probably will become as common for time-dependent processes as the straight line is for independent processes.[7] This model[8] may be described as $\hat{x}_{t+1} = \gamma_0 \hat{x}_t + e_t$.

One calculates a sum of squares SS of the deviations $(\hat{x}_i - x_i)^2$; $SS = \sum (\hat{x}_i - x_i)^2$ for any value of γ_0. Letting γ_0 take values from -1 to $+1$, one can plot SS as a function of γ_0, picking that value of γ_0 which minimizes SS.

Notice that, if there is an increasing trend in the series, the EWMA will always underestimate the series. One can correct this by modifying the model with a correction term called the cumulative or integral control: $\hat{x}_{t+1} = \gamma_0 \hat{x} + e_t - \gamma_1 (\sum_{i=0} e_i)$. That is, the predicted value of x_{t+1} equals the predicted value of x_t times a constant γ_0 plus the error (e) of prediction at t, minus a cumulative control parameter γ_1 times the sum of the previous errors. Table 1 illustrates two steps of an estimation process, using fictional data. With each successive step, the error is reduced.

Estimation of γ_0 and γ_1 proceeds by minimizing the residual sum of squares $SS = \sum_t (\hat{x}_t - x_t)^2$ with respect to both variables.

[7] Stuart Hunter, University of Wisconsin, personal communication, May 1968.

[8] The predicted value of x_{t+1} (written $\hat{x}_{t+1}$) equals the predicted value of x_t times a constant γ_0 plus the error (e) of prediction at t.

One can do this on a computer by having the computer plot SS as a function of γ_0 and γ_1 on a grid, and follow a search procedure (see Figure 5).

Box and Jenkins (1962) stated that they have rarely had to use the cumulative control parameter in industrial applications, but this modification of the EWMA model will make it sufficiently powerful for most purposes.

Reference to the statistician's experience with time series in industry for quality control and cybernetic control engineering is appropriate. The industrial problem involves charts of production output in a factory where (*a*) stability of production at an optimum level is to be maintained, and (*b*) changes in production are to be detected as a result of some administrative or technical intervention. For psychological problems, of course, behavior is usually the product.

Significance Testing

Suppose we wish to determine whether there has been a significant shift in the model fit for the series in Region I following a planned intervention, Event E (see Figure 6). Since the residuals are now uncorrelated—that is, merely white noise and hence independent—we can perform a simple F test.

We would merely assume that the model in Region I worked for Region II and proceed to

TABLE 1

AN ILLUSTRATION OF TWO SUCCESSIVE STEPS IN AN ESTIMATION PROCESS

		Time-series observations[a]			
Time	X[b]	Step 1		Step 2	
		$\hat{X}$[c]	Error	$\hat{X}$[c]	Error
0	1.000	1.000	.000	1.000	.000
1	1.000	.500	.500	.750	.250
2	1.000	.750	.250	1.125	.125
3	1.500	.720	.780	1.545	.005
4	2.000	1.140	.860	2.055	.055
5	2.000	1.430	.570	2.910	.910
6	2.000	1.285	.715	3.123	1.123
ΣE^2			2.897		2.311

Note.—In Step 1 $\gamma_0 = 0.500$; $\gamma_1 = 0.000$. In Step 2 $\gamma_0 = 0.500$; $\gamma_1 = 0.500$.
[a] Fictitious data.
[b] Actual.
[c] Predicted.

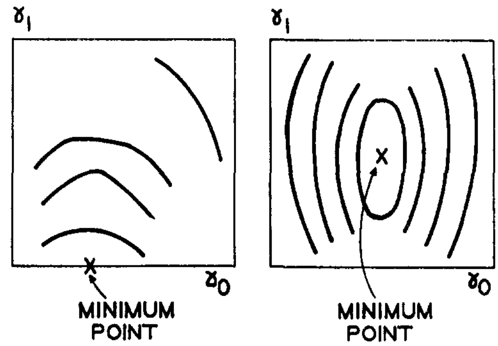

FIG. 5. Grid estimation of a two-parameter minimum (after Box & Jenkins, 1962).

forecast the observations in Region II, calculating a residual sum of squares $SS_1 = \sum (\hat{x}_t - x_t)^2$ where $\hat{x}_t$ is the forecasted value of x_t. Then we would fit Region II separately and calculate a residual SS_2. We then compute $F = \dfrac{(SS_1 - SS_2)/df_1}{(SS_1 + SS_2)/df_2}$, where $df_1 = 2$ because two parameters are involved in the estimation, and $df_2 = N_2 - 2$, where N_2 is the number of points in Region II.

To identify the causes of *unplanned* shifts in the model, one proceeds in a *post hoc* fashion to search for that point in the time series where the difference between the models for Regions I and II yields the maximum F. By consulting the log of concomitant events, hypotheses are formed regarding the cause of the shift. These hypotheses can then be tested by a replication or by a multiple time-series experiment.

For the analysis of a time-lagged multiple time series, as shown in Figure 7, one computes F as before: The generating function for Group 1 is derived and then is used to predict the behavior of Group 2 in the same region. The sum

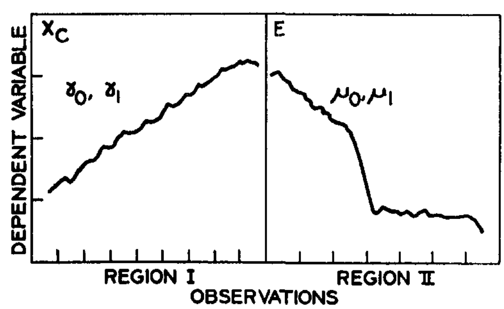

FIG. 6. Testing for shift in a time series.

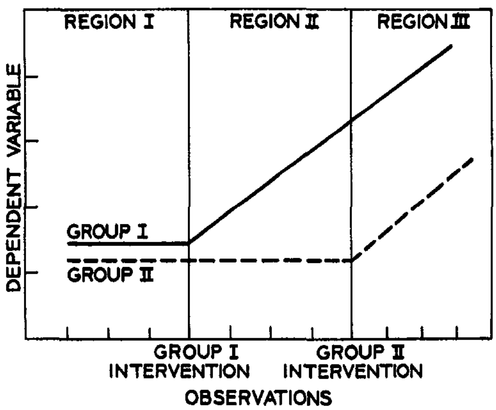

FIG. 7. A time-lagged multiple time series.

of squares of the differences between predicted and actual values equals SS_1. The SS_2 is the residual sum of squares of the generating function derived from Group 2 applied to Group 2. To assess treatment effects, one tests the hypothesis that Groups 1 and 2 differ only in Region II.

Application to Specific Problems

The question of whether the observation points in a time series should consist of group or individual data depends upon the nature of the problem being studied. When one is interested primarily in the pattern of group functioning or response to treatment and is not particularly interested in individual performance, a simple summation across subjects is appropriate. Here the generating function is applied to the group mean.

Analysis of variance procedures can use the average performance of a group over a period of time, but the group mean is not necessarily representative of the performance of any one subject. One of the unique contributions of the generating-function approach to time series is its power to assess this individual functioning. Once the performance of a single subject has been assessed, however, one normally wishes to generalize these findings to other subjects. There are at least three procedures for evaluating the generality of a single subject's time series.

First, each subject's time series can be considered a replication of the experiment. Using appropriate nonparametric tests, one may calculate the probability of obtaining the observed number of significant Fs.

Second, the generating function for the time series of one subject may be tested as a predictor of other subjects' time series. This procedure may involve a transformation. For example, if two heart-rate polygraphs are out of phase by half a period, a transformed generating function would result in successful prediction; simple summation would only obscure the obvious relationship between the curves.

Third, as a simple test, a directional correlation coefficient can be computed (Strahan, 1966). Wiener (1949) provided a sophisticated procedure for calculating a correlation coefficient between time series, providing information regarding both the signal and noise components of each subject's time series.

SUMMARY

The present paper has presented a research methodology which provides increments in precision and control appropriate to a number of research problems and settings. The core of this research methodology has been the time-series design, which has been presented as a powerful approach to be used in research settings where control groups are unavailable and/or where dependent observations are gathered over time (e.g., research in the field, or where $N = 1$). The time-series design has been shown to be a dynamic design, responsive to feedback in the sense that antecedent information can be used for subsequent planning and evaluation within an experiment. The choice of design, therefore, need no longer be a binary (yes-no) decision made in the initial planning phase of an experiment.

A mathematical model has been presented that utilizes the dependent nature of time-series observations. The advantages of this model over curve-fitting approaches were discussed. This model is appropriate to many research problems in the social sciences, and its implications have yet to be fully explored.

REFERENCES

ALEXANDER, H. W. A general test. *Psychological Bulletin*, 1946, 43, 533–557.

ANDERSON, R. L. Distribution of the serial correlation coefficient. *Annals of Mathematical Statistics*, 1942, 13, 1–13.

Bartlett, M. S. Some aspects of time correlation problems in regard to tests of significance. *Journal of the Royal Statistical Society, Series B*, 1935, **98**, 536–543.

Box, G. E., & Jenkins, G. M. Some aspects of adaptive optimization and control. *Journal of the Royal Statistical Society, Series B*, 1962, **24**, 297–343.

Box, G. E., Jenkins, G. M., & Bacon, D. W. Models for forecasting seasonal and non-seasonal time-series. In B. Harris (Ed.), *Spectral analysis of time-series*. New York: Wiley, 1967.

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

Campbell, D. T. From description to experimentation: Interpreting trends as quasi-experiments. In C. Harris (Ed.), *Problems in measuring change*. Madison: University of Wisconsin Press, 1967.

Campbell, D. T., & Stanley, J. Experimental and quasi-experimental designs for research and teaching. In N. L. Gaye (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally, 1963.

Coutie, G. A. Commentary on paper by Box and Jenkins. *Journal of the Royal Statistical Society, Series B*, 1962, **24**, 345–346.

Ezekiel, M., & Fox, K. *Methods of correlation and regression analysis: Linear and curvilinear*. New York: Wiley, 1966.

Flanagan, J. C. The critical incident technique. *Psychological Bulletin*, 1954, **51**, 327–358.

Grant, D. A. Analysis of variance tests in the analysis and comparison of curves. *Psychological Bulletin*, 1956, **53**, 141–154.

Holtzman, W. H. Statistical models for the study of change in the single case. In C. Harris (Ed.), *Problems in measuring change*. Madison: University of Wisconsin Press, 1967.

Mood, A. F. *Introduction to the theory of statistics*. New York: McGraw-Hill, 1950.

Strahan, R. F. A coefficient of directional correlation for time-series analyses. (Report No. PR-66-10 of The Research Laboratories) Minneapolis: University of Minnesota, Department of Psychiatry, 1966.

Watts, D. G. An introduction to linear modeling of stationary and non-stationary time-series. (Unpublished technical report No. 129) Madison: University of Wisconsin, Department of Statistics, 1967.

Webb, E., Campbell, D., Schwartz, R., & Sechrest, L. *Unobtrusive measures: Nonreactive research in the social sciences*. Chicago: Rand McNally, 1966.

Wiener, N. *Extrapolation, interpolation, and smoothing of stationary time-series, with engineering applications*. Cambridge, Mass.: M.I.T. Press, 1949.

Willems, E. P. An ecological orientation in psychology. *Merrill Palmer Quarterly of Behavior and Development*, 1965, **11**, 317–343.

Wold, H. *Bibliography on time-series and stochastic processes*. Cambridge, Mass.: M.I.T. Press, 1965.

Yule, G. V. Why do we sometimes get nonsense-correlations between time-series—A study in sampling and the nature of time-series. *Journal of the Royal Statistical Society, Series A*, 1926, **89**, 1–64.

(Received November 25, 1968)

ERRATA

In the article, "Short-Form Tests: A Methodological Review," by Philip Levy in the June 1968 issue, the citation of "Guerting, 1962" in the thirteenth line on p. 411 should read "Geuting, 1959," and the related reference on p. 415 should read as follows:

Geuting, M. P. Validities of abbreviated scales of the WISC. Unpublished master's thesis. Fordham University, 1959.