DOCUMENT RESUME

ED 103 382 SP 008 988

AUTHOR Schutz, Robert W.

TITLE Possible Solutions to the Problems of Measuring Change in Motor Behavior. Quantification Laboratory Technical Report No. 24.

PUB DATE Mar 75

NOTE 37p.; Paper presentel at the Measurement and Evaluation Symposium of the Annual Meetin of the American Alliance for Health, Physical E ucation, and Recreation (Atlantic City, New Jersey, March 1975)

EDRS PRICE MF-$0.76 HC-$1.95 PLUS POSTAGE

DESCRIPTORS Analysis of Variance; *Behavior Change; Criterion Referenced Tests; Factor Analysis; *Measurement Techniques; *Psychomotor Skills; *Statistical Analysis

ABSTRACT

This paper deals with problems of measuring change in motor behavior. Conventional measurement procedures and statistical analyses involving change are presented in the first section. This section discusses difference scores as the criterion measure and the use of all scores as the dependent variables. This latter category involves using either univariate or multivariate analysis of variance (ANOVA, MANOVA). The author gives suggestions for generally conservative researchers who want to use the methods discussed, but the second section describes alternative methods for analyzing change. The statistical techniques described are gradually being adopted by associated disciplines and are probably more appropriate for describing and predicting performance over time. This second section includes stochastic processes, time-series, factor analytic models of change, and curve-fitting as a change indicator. An appendix provides empirical comparisons among the numerous statistical methods described. (PB)

sol

POSSIBLE SOLUTIONS TO THE PROBLEMS of MEASURING CHANGE IN MOTOR BEHAVIOR

Presented
at the
Measurement and Evaluation Symposium
1975 National AAHPER Convention
Atlantic City, New Jersey
March, 1975

Robert W. Schutz
Quantification Laboratory
School of Physical Education
and Recreation
University of British Columbia
Vancouver, B. C.

Quantification Laboratory
Technical Report #24

SP009 088

"Nothing endures but change"
(Heraclitus, 500 B.C.)

"There is nothing in this world constant, but inconstancy"
(Jonathan Swift, 1707)

Change, the focus of many disciplines (history, geology, anthropology) and central to almost all scientific research, is such a necessary process and yet such a difficult one to measure. Science demands empiricism and, if we wish to infer causality, this empiricism must be in the form of controlled experimentation, often leading to a pre-post type of design with a resulting change or difference score. The problems inherent in the measurement of change have been well expounded (Dotson, 1973; Harris, 1963; Stelmach, 1975) but the solutions seem slow to develop. Bereiter (1963) has claimed that it is only in this area that he has heard of researchers abandoning research objectives due to the lack of suitable statistical procedures available. The task of providing valid solutions to the problems of measuring change is obviously a formidable one, and is certainly not going to be accomplished in this paper. What is presented, is a rather empirical account of the available methods for analyzing performance over time, the advantages and disadvantages of these procedures, and some biased, personal decisions regarding the "best" solutions. The discussion is dichotomized into two general approaches, one involving the common difference scores and repeated measures designs with their associated parametric statistical procedures, and the other approach focussing on alternate, less common, ways to study change, specifically; stochastic methods, time-series analysis, factor analytic procedures, and curve fitting.

A. CONVENTIONAL MEASUREMENT PROCEDURES AND STATISTICAL ANALYSES INVOLVING CHANGE

An overview of the educational and psychological literature dealing with the problems encountered in measuring change reflects that an indication of change is usually provided by two scores only - a pre-test and a post-test score, interspersed with some treatment condition or time lapse. However, research in sport and physical activity often results in a large number of responses per subject, rather than just a pair of scores, thus allowing for a greater variety of possible designs and analyses. Consequently, it is necessary to examine the conventional measurement of change as two distinct processes, one involving a difference score, the other utilizing all the data in a repeated measures design.

1. Difference Scores as the Criterion Measure.

a) Selection of a Criterion Score (unadjusted).

If the research methodology utilized yields a single score on the first administration of a test (X_1) and another single score on a repetition of that test at some subsequent point in time (X_2), then there is little choice in the criterion score to use if the researcher wishes to use a single, unadjusted, dependent variable. It has to be this difference ($D = X_2 - X_1$) which has many inherent deficiencies and numerous possible transformations to reduce these deficiencies (none of which are very satisfactory). These are discussed in section 1(b). A more likely situation, however, is when there are a number of observations available for each S (e.g., heart rate at each minute of a 15-minute exercise bout, 30 learning trials), but the investigator wishes to reduce this data to a single change score or learning score. The problems then confronting him are: (1) how many trials should he use to estimate both the initial and final states of the Ss?, and (2) should he use the best, or the average, of each of these sets of trials? Before commenting on some possible solutions to these two problems, it should be noted that neither of these problems should ever arise when dealing with the analysis of change. Discarding or reducing data, when suitable statistical methods are available for analyzing all available data, seams like very inefficient research. If the goal is the be able to understand motor behavior, for purposes of explanation and prediction, then one must look at all the data, and analyze it by a repeated measures ANOVA, time series, or some other equally suitable tool. However, many investigators insist on obtaining a single change score, thus some discussion on these points seems necessary.

The problem of choosing between the best and the average score has only one acceptable solution - use the average. There is sufficient support for use of the average rather than the best in the general case (Baumgartner, 1974; Henry, 1967; Kroll, 1967) and in the specific case of difference scores it is even more necessary. The reliability of a difference score is so dependent upon the reliability of the two scores which produce this difference, that it is imperative that these two scores possess maximum reliability themselves - thus averages are necessary.

The solution to the question of the optimal number of trials to use in computing these pre and post-score averages is not quite so unambiguous. The problem facing an investigator who uses a learning task is how can he choose a score which maximizes both reliability and discriminability at the same time? In a task which has, say, 20 trials, the difference between trial one and trial 20 will probably show the greatest discriminability as far as learning is concerned; however, it may not be very reliable. If one uses the average of the first ten trials as an indication of initial score, and the average of the last ten as the performance score, then the difference between these two may show high reliability, but it probably will not show much learning. Carron and Marteniuk (1970) pointed out the necessity for comparing the differences between both the reliabilities and discriminability obtained by grouping trials in different ways. Others (Baumgartner and Jackson, 1970; McCraw and McClenney, 1965) have attempted to give definitive rules for determining the number of trials and the measurement schedules one should employ. Because of the great variability in type of task, characteristics of Ss, etc., it does not seem possible to choose a specific rule for determining the "best" criterion measure for all situations - even for all situations involving a specific task or set of measures. If one decides that it is necessary to reduce the data to a single dependent variable (which, to this writer, does not seem to be a valid procedures), then utilizing procedures as suggested by Carron and Marteniuk (1970), and following the basic principles of reliability and validity of dependent variable scores which have been frequently and explicitly laid out for us (e.g.; Alexander, 1947; Burt, 1955; Feldt and McKee, 1957; Krause, 1969; Lomnicki, 1973; Schutz and Roy, 1973) one should be able to arrive at a procedure for selecting the most suitable criterion score in each specific situation.

b) Selection of a Criterion Score (adjusted).

In situations where there are only two opportunities for observation and measurement (pre and post), or where the investigator insists on reducing repeated measures to a pre-post case, then it is probably necessary to apply some type of statistical adjustment or correction factor to either the difference score or to the final score. The following section gives possible solutions for each of a number of common problems associated with using difference scores.

These problems have been well defined by many investigators (Bereiter, 1963; Cronbach and Furby, 1970; Lord, 1956, 1963; McNemar, 1958).

(i) Problem 1. Regression Effect: In general, on the second administration of a test, and in the absence of any true change or treatment effect, the observed scores for those who scored high on test #1 tend to decline and the observed scores of those who scored lowest on test #1 tend to increase on test #2.

Solutions. The most valid, and least complicated, solution, is to use a homogeneous group so all Ss have essentially the same initial score. If the experiment involves comparisons between groups, then equate the group means initially, either by randomization with large sample sizes, blocking, matching, or statistically through analysis of covariance (these methods are discussed below in Section II(a).

Another possible solution, the one to which psychometricians have directed their attention, is to adjust the final score on the basis of the pre-post linear regression effect. This can be done by fitting a regression line to the pre-post scores (X_1, X_2) under the conditions of the null hypothesis; i.e., no treatment effect, and then use deviation from the regression line as the dependent variable indicating true change (Lord, 1963). This requires either a separate control group or a (X_1, X_2) measure for each subject under a treatment condition and a control condition - a procedure which is not always possible. The most reasonable solution seems to be to use analysis of covariance (ANCOVA) as it is essentially an analysis of the X_2 scores, adjusted on the basis of the regression line between X_2 and X_1.

(ii) Problem 2. Measurement Errors or the Unreliability-Invalidity Dilemma: The degree to which measurement errors exist in the initial and/or final measures, along with the degree to which the X_1, X_2 correlation exceeds zero, is reflected by a reduction in the reliability of the X_1-X_2 difference score.

Solutions. There exists a wealth of information on possible solutions to this problem (e.g.; Lord, 1956, 1963; McNemar, 1958; Ng, 1974; Tucker, 1966; Wiley and Wiley, 1974).

The basic thesis of all these articles is that it is possible to compute a reliability coefficient 'corrected for attenuation', that is, the reliability of a difference between 'true scores' (errorless measures yielding reliabilities of 1.00 in both X_1 and in X_2). Once having obtained a reliable estimate of true difference it is then possible to use this attenuated reliability coefficient and multiply it by the observed X_2-X_1 difference (but scaled as deviations from the means), thus obtaining a hypothetical true difference score or "regressed score" (McNemar, 1958). Although this is the basis of the solutions advocated by many psychometricians it has its deficiencies, the primary one being that the number of alternate ways to compute this true gain score seems to be exceeded only by the number of papers written on the topic. The non-specialist is left with a morass of equations and confusion. Another deficiency with the use of estimated true difference scores is that the regression coefficient used in the predictor equation is based on a number of assumptions, some of which may not always hold true. A recent report by Wiley and Wiley (1974) indicates that the assumption of independence of errors of measurement between tests is frequently violated, thus giving overestimates of the attenuated reliability coefficient. This in turn would result in overestimates of the true gain score.

(iii) Problem 3. Equality of Scale Along the Range of Scores (the Physicalism-Subjectivism Dilemma): An observed score at the low range of the continuum may be measuring an attribute of behavior quite different from that which is reflected by the same test at the high end of the range of scores.

Solutions: There seem to be no adequate solutions per se for this problem. One could use P-technique methodology (a sort of factor analysis appropriate for change data) to test the assumption that the two measures are in fact measuring the same thing (Bereiter, 1963; Cattell, 1963). However, this is not a solution, but rather a technique to reveal the existence or non-existence of a problem. The answer seems to be in finding ways to avoid the problem rather than solve it - and this can be accomplished to a limited degree. If all groups are equated initially with respect to their scores on the dependent variable, then any differences between groups in the amount of change within groups can be logically interpreted (Schmidt, 1972).

This restriction allows for the conclusion that one group changed more, or less, with regards to the particular dependent variable being used. If one group showed very large changes, and the other group very small ones, then it may be difficult to interpret the meaning of the relative magnitudes of change scores, but it is still possible to state that one group should significantly greater change than the other group on that particular trait.

A General Solution to the Problems Associated with Difference Scores:

At this point the reader must be wondering, "Is there no adequate solution to the problem of measuring change?" My answer is "Yes" there are adequate methods, but not through the use of difference scores. If one must use a change score, then perhaps the "best" estimator of a true difference score is Cronbach and Furby's "complete estimator" (1970):

$$\hat{D}_{\infty} = \hat{X}_{1_{\infty}} - \hat{X}_{2_{\infty}}$$

where $\hat{D}_{\infty}$ is the "true difference score", and $\hat{X}_{1_{\infty}}$ is the true score at time 1, taking into account numerous other categories of variables, W, which may be multivariate in nature and relate to the pre or post scores in some manner. The true score for X_1 is estimated as:

$$\hat{X}_{\infty} = p_{xx'} X_1 + \frac{\sigma X_{1_{\infty}}(X_2 \cdot X_1)}{\sigma^2(X_2 \cdot X_1)} (X_2 \cdot X_1) + \frac{\sigma X_{1_{\infty}}(W \cdot X_1, X_2)}{\sigma^2(W \cdot X_1, X_2)} (W \cdot X_1, X_2) + \text{constant}$$

where $(X_2 \cdot X_1)$ and $(W \cdot X_1, Y_1)$ are partial variates. The purpose of presenting this equation is not to provide the reader with a useful statistical tool, but rather to point out the extreme degree to which the raw data can be transformed if one wishes some sort of pure measure. The difficulty in interpreting this transformed score is obvious - at least in terms of predictable observed behavior.

Two quotes provide a suitable summary of this investigator's position on the use of difference scores:

> "Both the history of the problem and the logic of investigation indicate that the last thing one wants to do is think in terms of or compute such change scores unless the problem makes it absolutely necessary." (Nunnally, 1973, p. 87)

> "Gain scores are rarely useful, no matter how they may be adjusted or refined." (Cronbach and Furby, 1970, p. 68)

2. The Use of All Scores as the Dependent Variables.

The analysis of all of the available data should provide an investigator with more information than does the limited, and suspect, information provided in a difference score. These repeated measures analyses may be performed by either univariate or multivariate analysis of variance (ANOVA, MANOVA) on the raw scores or on scores adjusted for initial differences between groups. The more information available on the nature of change in behavior over time, the greater should be the degree of understanding of the nature and causes of that change. Consequently, in an experiment involving any length of time between the initiation of the treatment and the final observation, it is desirable to take numerous measures per S. Although in some cases it is not possible to do this, either due to the contamination effect of the measurement tool or to the nature of the treatment procedures, in most motor behavior studies such repeated measures are quite feasible.

a) Repeated Measures ANOVA.

The common method for analyzing change for a repeated measures design is through a repeated measures or Ss x Treatments ANOVA. Given a typical experiment involving two treatment groups (or a treatment and control) with 20 Ss nested within each group and repeated across say 10 trials (Fig. 1), one appropriate method for analyzing change could be to break down the total variability as given in Table I.

[Insert Fig. 1 and Table I about here]

The effects of most interest here, with respect to the analysis of change, are the Groups x Trials and its trend analysis components, Groups x Trials (Linear) and Groups x Trials (Quadratic). The Groups x Trials interaction indicates the degree to which the change over trials is the same for each group - which is probably the research question of most interest; i.e., is there a significant change in behavior over the time span of the experiment, and, if so, does this change show the same, or different, characteristics between the two experimental groups?

		Trials			
		T_1	T_2	. . .	T_{10}
Group 1	S_1	X_{111}	X_{112}	. . .	X_{1110}
	S_2	X_{121}	X_{122}	. . .	X_{1210}
	.	.	.		.
	.	.	.		.
	.	.	.		.
	S_{20}	X_{1201}	X_{1202}	. . .	X_{12010}
Group 2	S_1	X_{211}	X_{212}	. . .	X_{2110}
	S_2	.	.		.
	.	.	.		.
	.	.	.		.
	.	.	.		.
	S_{20}	.	.		X_{22010}

Fig. 1. Schemata of 2 x 10 Factorial Experiment with Repeated Measures on the Second Factor.

TABLE 1

Analysis of Variance, with Trend, for a 2 x 10 Factorial Experiment with Repeated Measures on the Second Factor

Source	df	Mean Square	F Ratio
Groups	1	MS_G	$MS_G/MS_{S(G)}$
Ss within Groups	38	$MS_{S(G)}$	
Trials	9	MS_T	$MS_T/MS_{S(G)T}$
Linear	1	MS_{T_L}	$MS_{T_L}/MS_{S(G)T}$
Quadratic	1	MS_{T_Q}	$MS_{T_Q}/MS_{S(G)T}$
Residual	7	MS_{T_R}	$MS_{T_R}/MS_{S(G)T}$
Groups x Trials	9	MS_{GT}	$MS_{GT}/MS_{S(G)T}$
G x $T_{Lin.}$	1	MS_{GT_L}	$MS_{GT_L}/MS_{S(G)T}$
G x $T_{Quad.}$	1	MS_{GT_Q}	$MS_{GT_Q}/MS_{S(G)T}$
G x $T_{Resid.}$	7	MS_{GT_R}	$MS_{GT_R}/MS_{S(G)T}$
SwG x Trials	342	$MS_{S(G)T}$	
Total	399		

The Groups x Trials (Linear) asks essentially the same question but with the constraint that the change over time is linear. In this case a linear function is forced on the data and the test of significance tests for equality of slopes between the two groups, which in behavioral terms amounts to a comparison of the rates of learning, rates of recovery, etc. Similarly the Groups x Trials (Quadratic) compares the two treatment groups on the basis of the degree of curvature or time of plateauing of the scores over time.

This analysis then provides one possible solution for the analysis of change suitable for many experimental conditions. By using a number of measures instead of just two, the problems of regression effect and measurement errors are greatly reduced. The unreliability of the data is reflected by the magnitude of the S x Trials interaction (or in this case the S(G) x T) and is thus a sort of built in protection against making erroneous research conclusions based on unreliable data. The less reliable the data is, the larger the S x Trials error term, the more difficult it is to attain statistical significance and the less likely it is to make a Type I error.

The repeated measures ANOVA is not the ideal solution to the problems of analyzing change, however, for a number of reasons. Firstly, the tests of significance give limited information regarding the nature or form of the change over time, as the trend analyses fit only polynomials to the data, data which is frequently better fitted by a logarithmic or exponential function. Secondly, it deals with mean values only and does not reveal reliable differences between subjects (within the same group) with respect in intraindividual b~havioral changes over time (a stochastic model would detect this). Finally, and perhaps most importantly, the nature of the data common to most studies in motor behavior is such that it violates the assumptions on which the repeated measures ANOVA is founded. These assumptions are that the measures (i) are normally distributed, (ii) exhibit equal variances under all treatment conditions, and (iii) have equal covariances between all treatment pairs (the precise mathematical assumption is that all covariances equal zero but the F ratio is virtually unaffected by violation of this assumption, providing all covariances are equal). While the first two of these assumptions are usually met with motor performance data, the third one rarely is.

This assumption can be casually tested by examining the correlation matrix of the repeated measures - the degree to which all correlations are not equal indicates the degree to which this assumption is violated.[1] It is frequently the case in our field of study to obtain data in which adjacent trial correlations are very high, but diminish as a function the number of intervening observations between any two measures. The resultant of this situation is an inflated F value and a substantial increase in the probability of committing a Type I error (as high as $p = .15$ when assuming a $p = .05$).

The analysis of variance for repeated measures, which was first presented here as a possible solution to some of the problems inherent in the analysis of change, has now become a problem itself. There are two possible ways by which ANOVA may be validly used on repeated measures data which exhibits unequal between trial correlations:

(1) Inflate the magnitude of the F needed for significance by reducing the associated degrees of freedom (d.f.). Box (1954) has suggested that the d.f. for both the numerator and denominator be multiplied by a factor ε, which is a function of the degree of heterogeneity of both the variances and the covariances. The greater the heterogeneity the smaller the calculated ε and the larger the F value must be in order to reject the null hypotheses.

(2) Greenhouse and Geisser (1959) questioned the validity of the estimator ε and its effect on the approximate F distribution. They suggested the use of the minimum possible value of ε, namely $1/(k-1)$ where k is the number of levels of the repeated factor, as the factor which should be applied to the d.f. in all situations. Although this is a statistically valid technique it is very conservative, thus resulting in a rather large probability of committing a Type II error.

There are a number of excellent articles available which provide a lucid explanation of both the problem and the merits of these solutions (e.g., Davidson, 1972; Gaito, 1973; Gaito and Wiley, 1963; McCall and Appelbaum, 1973; Mendoza, Toothaker and Nicewander, 1974).

[1] Procedures for statistical tests of this assumption are available in Winer (1971, p. 594).

b) Repeated Meausres MANOVA.

The other solution to the problem of non-homogeneity of covariances is to use a technique which does not require this assumption - namely the multivariate analysis of variance. MANOVA requires no assumptions regarding the homogeneity of covariances and allows for an exact statistical test based on a known significance level. Although this technique has been available for many years, it has not been adopted by practicing researchers due to its extreme computational complexity. However, the present accessibility of suitable computerized multivariate statistical packages at most universities has eliminated such an excuse for ignoring this very useful test and it should now be a standard statistical tool for all researchers. Very briefly, what MANOVA does is to transform the k repeated measures for each subject into a set of (k-1) scores through the application of independent contrasts (these are usually orthogonal polynomials, but they need not be as the resulting significance test is independent of the choice of contrasts). An analysis of variance type procedure is then carried out on the vector of means of these derived scores with the mean square error being a variance-covariance matrix of within cell variabilities rather than a unitary scalar value as in the univariate procedure. The tests of significance provide an F ratio for the overall multivariate hypothesis that the trial means are equal, and for a two group experiment, that the change in performance across repeated measures is the same for each group. An overall significant F on these multivariate hypotheses allows the investigator to use appropriate follow-up tests while maintaining an overall pre-determined level of significance. These follow-up procedures can take the form of simultaneous confidence intervals, step-down F ratios, or even the usual univariate F tests on each dependent variable separately or on the single d.f. contrasts associated with trend analysis.

Another frequently used procedure associated with MANOVA is discriminant analysis which tests whether two or more groups can be significantly separated on the bases of their profiles (or, in the RM design, their pattern of change over time). It has been shown, however, that a Groups x Trials ANOVA is more versatile in detecting the nature of the differences between group profiles than is discriminant analysis (Thomas and Chissom, 1973). Although Thomas and Chissom failed to consider the restrictive assumption inherent in the univariate G x T ANOVA, this is not a factor if the Trials effect is broken down into polynomial coefficients (linear, quadratic, etc.).

This essentially converts the univariate procedure to a multivariate technique and thus no longer requires the assumption of equal covariances. Bock (1963), Cole and Grizzle (1966), and Finn (1969) have provided comprehensive discussions on the application of MANOVA to repeated measures data, and comparisons of the applications and outcomes of ANOVA versus MANOVA are well given by Davidson (1972), Hummel and Sligo (1971), McCall and Appelbaum (1973), and Poor (1973).

c) Experimental and Statistical Adjustments for ANOVA and MANOVA.

As was stated above, a number of the problems associated with the measurement of change can be reduced if all treatment groups are initially equal with respect to the dependent variables. The four procedures available for achieving this initial equality are: random assignment, balancing or matching, blocking and analysis of covariance.

(i) Random Assignment. This, theoretically, is the best way as it equates groups initially with respect to all variables. Unfortunately, the success of random assignment is dependent upon the size of the samples and the population variability of the independent variable of interest. Samples of size 100 almost guaranty equality (but it is never a certainty), whereas samples of size 5 are rather unlikely to result in equal distributions among the treatment groups.

Some investigators advocate the random assignment of Ss to treatment groups followed by a t test (sometimes with an exaggerated alpha, say .90) to determine if the hypotheses of initial equality can be accepted (Rosemier, 1968). If the hypotheses of initial equality is not tenable, then the investigator needs to either reassign Ss to treatments, increase his sample size in the hope that the randomization process will eventually work, or adjust his groups with some type of balancing procedure. None of these procedures are very satisfactory, from a statistical as well as a procedural aspect.

(ii) Balancing and/or Matching. These procedures involve assigning Ss to treatments on the basis of their initial scores (or some other related variable) in an attempt to equate groups initially. It has been shown that matching is always less efficient than analysis of covariance and is usually less efficient than simple random sampling (Billewicz, 1965).

It has also been shown (Finney, 1957) that matching is never as suitable as blocking. The obvious recommendation is that these procedures should not be used anymore in our empirical research studies.

(iii) Blocking. Blocking, when done on the basis of initial scores, is essentially the same idea as balancing; however, the sampling and assignment procedures are quite different, thus making blocking a statistically sound procedure. Correct blocking technique requires knowledge of the distribution of the blocking variable in the population, an *a priori* determination of the cut-off values which determine the blocking levels, and then sampling from each of these population strata to form the blocks.

(iv) Analysis of Covariance. Whereas blocking provides an experimental method of equating groups, analysis of covariance provides a statistical method for doing so. The choice between these two techniques is not a simple one, as the relative advantages of one procedure over the other depend upon the degree of relationship between the concomitant variable (used for blocking or as the covariable) and the dependent variable. Feldt (1958) has shown that if the correlation between the concomitant variable and dependent variable is less than .60, blocking is better, whereas if it is greater than .80, analysis of covariance provides a more powerful statistical test. However, Feldt suggests that even with high correlations, blocking is preferable as the relatively small advantage in precision shown by analysis of covariance is more than lost due to the strict assumptions of regression inherent in covariance; i.e., linearity of regression, and equality of regression within treatment groups.

A Summary of Section A

For those generally conservative researchers who wish to restrict their statistical analyses to conventional parametric techniques, here are some guidelines:

1. Use MANOVA - with trend analysis and covariance if necessary.
   a) Obtain a series of measures on each subject throughout the treatment period when change is expected.
   b) Use at least 20 more subjects than there are measures per subject.
   c) Test for equality of groups initially - if they are not equal, then use the initial score as a covariate.
   d) Analyze the data using MANOVA procedures.

2. If difference scores are required for experimental or theoretical reasons, then make the best of them by:
   a) Attempt to maximize the reliabilities of the pre-test and post-test scores and minimize the pre-post correlation (while at the same time maintaining equality of meaninging between the two sets of scores).
   b) Equate the groups initially as best as possible - either through randomization with a large N, or through blocking on relevant variables.
   c) Compute the reliabilities of the difference scores so the data may be interpreted with the required caution.
   d) Analyze the data with a t-test or ANOVA.

B. ALTERNATE METHODS FOR ANALYZING CHANGE

Within the discipline of human kinetics, the usual methods for analyzing change are the deterministic, parametric methods discussed above. However, there are a number of alternate statistical techniques gradually being adopted by associated disciplines which may not be as precise in terms of hypothesis testing, but are probably more appropriate for describing and predicting performance over time. Included in these procedures are three techniques which have potential as useful statistical tools for the analysis of change of motor performance data; namely, stochastic methods and the associated time series analyses, factor analytic techniques for measuring change, and curve-fitting.

1. Stochastic Processes

A stochastic variable, which may be defined as a time-dependent variable, refers to any dependent measure which is observed repeatedly over time. Thus, in this content, all change scores are stochastic to some extent. A more general use of the term stochastic, however, is through its association with stochastic processes and Markov chains - a series of time dependent events which are related to each other by a transition probability. A transition matrix, composed of a number of these transition probabilities, defines the probability that a dependent variable or measure will make a change (of some specified magnitude) during the time between two successive observations.

Stochastic processes are usually described in terms of a set of discrete states and a set of one-step transition probabilities.

The states are classifications of the variables under observation, such as the number of errors made in a learning task, the attitude of an individual toward physical activity at a certain point in time (e.g., favorable, indifferent, unfavorable), or a heart rate at various stages of activity (altered from a ratio scale to an ordinal scale or nominal classification). In general stochastic processes can be divided into four distinct classes: (a) discrete state - discrete time, (b) discrete state - continuous time, (c) continuous state - discrete time, and (d) continuous state - continuous time. Type (a) is the process most commonly applied to models in the behavioral sciences as measurement, calculation, and interpretation all become more difficult in the continuous cases. Queueing processes, and the birth-death processes of ecology and genetics are examples of the second type. The third and fourth types of stochastic processes are less commonly used (see Bailey, 1964; or Karlin, 1966, for examples).

The statistical analysis of change data through the application of stochastic models will yield both descriptive and inferential statistics which can help the researcher test his theories. Descriptive statistics of interest are such values as: the transition probabilities themselves (and comparisons among transition probabilities under different experimental conditions); the asymptotic value of a transition probability from time one to some very distant time; the expected number of trials before learning, fatigue or some such absorption state occurs; and the probability of being in some particular state at a specified point in time. These statistics, which are calculated directly from the observed data, can then be compared with theoretical values calculated from the theorems of a model. Such comparisons prove very helpful in isolating faulty assumptions in the theory. For example, it could happen that the observed values for the total number of errors and the number of times the process was in a particular state both agreed closely with the theoretical values, but the observed variance of the mean number of errors deviated substantially from the theoretical value. This would suggest that perhaps a more realistic model could be developed by using, say, a four state process rather than the two or three state one originally hypothesized.

The application of inferential statistics requires knowing the distribution of the particular test statistics before any probability statements can be made.

Such distributions have been established by Anderson and Goodman (1957) for making statistical inferences about Markov chains (a stochastic process in which the probability transition from one state to another is dependent only upon the state of the process at the previous time). Knowing these distributions (they are all asymptotically distributed as χ^2 with various degrees of freedom) the following null hypotheses may be tested.

a) The transition probability is independent of t; that is, test the stationarity of the process to see if the transition from one state to another is the same no matter what trial it occurs on.

b) The stochastic process defined by treatment group one is the same Markov chain as the process defined by treatment group two. If this hypothesis is rejected and in fact the group one data fits a first-order chain and group two a second-order chain, this tells the researcher that the trial to trial scores for group two exhibit a greater degree of dependency on the past than do the scores for group one.

c) In a process involving two sets of states, the transition probabilities in one set are independent of those in the other set of states. For example, the two state spaces may be levels of respiratory rate and levels of heart rate during continuous exercise. This hypothesis tests whether the sequence of changes in respiratory rate is independent of the sequence of changes in heart rate. This is not at all the same as the usual hypotheses which tests whether a series of discrete respiratory rates are independent of a series of discrete heart rates.

Once the statistics as predicted by the model have been compared with the ones calculated from the observed data the investigator has a good indication of the adequacy of his model. More specifically, if the data do not agree with the model, he can tell exactly where the model and the data were incompatible and made the necessary adjustments to the appropriate theorems or assumptions of the model. Barring a very gross misrepresentation of the data by the model it is not necessary to discard the whole theory. In general, lack of agreement between the model and the observed data may be due to one or more of the following: inappropriateness of the model (the model requires a change in theorems or assumptions), errors in the design and execution of the experiment (perhaps better experimental controls will eliminate the effect of some extraneous variables), or a flaw in the theory upon which the model was based (the model-observation discrepancy should suggest the appropriate theoretical revisions).

Stochastic methods have been used rather extensively in psychology, primarily in the area of learning, (see, for example, Greeno and Bjork, 1973, who list 243 references dealing with mathematical learning theory, a large number of which are stochastic in nature) and to a lesser extent in sociology (Carlson, 1972; Guppy and Fraser, 1973). In the area of sport and physical activity we are just beginning to examine the possibilities of stochastic methods, but so far have very little empirical support of its usefulness over the more conventional statistical techniques. Schutz (1970a) has described its potential on a theoretical basis, and provided an example of its practicality as an analytical tool in evaluating scoring systems (1970b). However, he has also provided an interesting example of how a behavioral theory can be represented by a rather complex stochastic model which leads to nothing but confusion and mathematical merry-go-rounds (Schutz, 1971). Guppy and Fraser (1973), by using a Markov model to examine occupational mobility in professional sport, showed that baseball players have differential mobility rates according to race. Other ongoing research (by Rennick at the University of Washington and Salmela at the University of Laval) may provide us with further examples of the advantages of stochastic processes, but until such time as a number of published research articles appear which clearly show that stochastic methods provide greater insight into the interpretation of empirical data than do standard statistical procedures, their general adoption cannot be recommended.

2. Time-Series

A time-series experiment involves repeated measures on one or more individuals over a period of time, thus the resultant observations for each individual are time dependent and usually correlated (in effect, stochastic). Under these conditions repeated measures ANOVA procedures are not appropriate, and, unless the sample size is large relative to the number of observations per individual, neither is MANOVA. Methods for analyzing data from time-series experiments, which may be done on repeated measures from a single individual or on the means of a number of individuals, have developed rather recently and consequently have not yet been used extensively in empirical research. Statistical models for testing the significance of the change in level of a nonstationary time-series and for comparing time-series among different treatment groups have been proposed by a number of statisticians in the past ten years (Box and Tiao, 1965; Jones, Crawell and Kapuniai, 1970; Glass and Maguire, 1968; Gottman, McFall and Barnett, 1969; Shumway, 1970; Strahan, 1971).

Most of these methods involve rather complex matrix manipulations and for this reason, along with the fact that they have not been used to any extent in empirical research studies, they are not recommended to the non-statistician in our field at this time. One time-series procedure which has been shown to be useful, however, is the autocorrelation (serial correlation) which provides an indication of the degree of sequential dependencies among the successive observations. A serial correlation of lag one (r_1) is obtained by pairing the first observation with the second, the second with the third, etc., and then calculating the product-moment correlation coefficient on these n-1 pairs (n being the number of repeated observations). Similarily, serial correlations of lag 2, 3, etc., (r_2, r_3) can be calculated. Each coefficient by itself gives some information on the serial dependencies in the data, and if one were to plot r_t against t (the time lag) for successively increasing values of t, the resultant graph, or correlogram, would indicate the change in serial dependencies throughout the total series.

Correlograms are particularly useful for experimental situations in which a series of repeated measures are obtained before and after a treatment is administered. A change in the nature or degree of the serial dependencies following administration of a treatment indicates a significant treatment effect (the test for statistical significance of a serial correlation coefficient is the same as that for an ordinary product-moment correlation coefficient). Other appropriate situations for utilizing a time-series are the two-group case in which the correlograms of the two groups can be compared, and experiments involving measures on a number of dependent variables at each point in time. This latter case lends itself to multiple time-series analysis, involving cross correlations (serial) between the variables, and thus tests the extent to which trial-to-trial variation in one variable can be attributed to concomitant trial-to-trial variation in another variable (Holtzman, 1963).

3. Factor Analytic Models of Change

The use of various factor analytic methodologies as a powerful tool for analyzing change has been advocated by a number of researchers, especially those in developmental psychology (e.g.; Baltes and Nesselroade, 1973; Bentler, 1973) and educational psychology (Corballis, 1970; Harris, 1963). These procedures, requiring multiple measures at each point in time, may involve the comparison of factor loadings and factor scores between time periods (Corballis, 1970), or may require an extension of the usual two-way data matrix (subjects by variables) to a three-way data matrix (the third variable being occasions) and its resultant, and rather complex, factor structure (Tucker, 1963). In motor behavior research we usually restrict our dependent variables to a few (five or less), and thus factor analytic procedures are not appropriate. For this reason (along with the fact that this investigator has had no previous experience with factor analytic change models), and also because it is not appropriate for the data used for the empirical examples given in this paper, no further discussion of factor analysis and its associated image and canonical analyses are presented here. Readers interested in this method are encouraged to read the references previously mentioned and attempt to apply these methods to motor behavior data. Unfortunately we are all somewhat reluctant to attempt a new technique until we are provided with empirical evidence that it will tell us something that the conventional, established procedures do not. Factor analytic change models may be useful tools - we need someone to prove this to us.

4. Curve-Fitting as a Change Indicator

The fitting of a mathematical function to a set of points spaced along a time continuum can be done in a number of ways, the most common of which is the previously mentioned trend analysis. Trend analysis will fit a set of orthogonal polynomial coefficients to a series of trial means, yielding an F ratio for each degree polynomial. This provides as estimate of the degree of linearity, quadratic curvature, etc., displayed by a series of points, or, if there is more than one treatment group, the groups x trials (linear), etc., effects indicate the difference between groups in the nature of the change in performance over the total series of trials. While this procedure is adequate if the data are indeed of a polynomial nature, two problems arise if it is not. Firstly, it is obvious that if the data can be better represented by an exponential or logarithmic function, then the best fitting polynomial is less than adequate.

The second problem associated with using trend analysis is that the curve is fitted to the trial means rather than to the individual scores of each trial for each subject. If the data is in fact polynomial in nature, then the function fitting the trial means will be identical to that obtained by finding the function for each subject and then averaging the coefficients of the individual equations. However, if the data is exponential, then the curve of the means may grossly distort the typical individual curve in that it will smooth out reliable and consistent discontinuities in the data. This has been clearly shown by Merrill (1931) with growth data, and by Sidman (1952) with learning scores.

Examples of fitting non-polynomial curves to motor behavior data are not uncommon; however, in most cases the investigators restrict their analyses to descriptive techniques only, that is, they find the best fitting function, and attempt to interpret it in a subjective manner. Furthermore, it seems that at times researchers attempt to find the best fitting mathematical function - without regard to the theoretical meaning associated with the parameters of the derived function. While it is true that such a function may be useful for predictive purposes, it is of little value in the description and explanation of behavior.

The procedures recommended here for analyzing change through the application of curve-fitting are as follows:

a) Select, a priori, the type of function which best represents the underlying physiological or psychological process hypothesized. The function should be simple enough so that the parameters are interpretable and can differentiate between treatment groups. For example, the exponential function

$$y = a + be^{-ct}$$

represents a negatively decelerating function suitable for a number of motor performance data sets. The parameter a reflects the asymptotic value of y (its minimum in this case, which will be reached eventually), the parameter b indicates the total change in y from time zero to asymptote, and c describes the rate of change in y with respect to time t.

b) After collecting the data, fit this function to the series of data points for each subject. Thus each subject now has three dependent variable scores, a value for each of a, b, and c.

c) Determine the percent of variance accounted for by the function and either accept it or reject on the basis of an a priori cut-off level.

d) Assuming that there are two or more treatment groups, ANOVA can now be performed on each of the three dependent variables, providing tests of hypotheses on differences among the groups with respect to: asymptotic performance, total amount of change, and rate of change.

A general theoretical explanation of these procedures, along with suggestions for more sophisticated techniques, is provided by Snee (1972), and Henry and DeMoor's (1950) article gives an excellent example of this type of methodology.

APPENDIX

The following tables and figure are provided for empirical comparisons among the numerous statistical methods suggested in this paper. Three sets of data were computer generated, each one simulating a 2 x 20 factorial design with repeated measures on the second factor. Factor one represents two treatment groups (n = 30/group), and Factor two can be considered a days or trials factor. The three experiments represent different conditions of the variance-covariance matrix, but all had the same means and variances (a constant variance of 10.0 for all trials, and means ranging from 7.0 to 31.0). Fig. 2 shows the trial means for each treatment group, and Table 1 gives the exact values for each case. The three cases representing different covariance structures are:

Case 1: A constant covariance of 2.0 between all pairs of trials, thus yielding an $r_{ij} = .2$ for all trials i,j $(i,j = 1, --- 20; i \neq j)$

Case 2: A constant covariance of 8.0 between all pairs of trials $(r = .8)$.

Case 3: A varying covariance, ranging from 9.0 for adjacent trials to 1.0 for trials 15 or more steps apart $(r = .9 \text{ to } .1)$.

Tables 2 and 3 give the F ratios and error variances for each of the statistical tests commonly used to analyze change. The t tests at the top of Table 2 are included as this procedure is used occasionally, even though it is completely invalid. In this 2-t-test procedure a t value is calculated on the difference (Post test - Pre test) for group I and another t for group II. A subjective assessment is then made on the relative magnitude of the two t's. The other tests are standard statistical procedures using various forms of the dependent variable (difference scores, trial 20 minus trial 1; difference scores, mean of trials 18-20 minus mean of trials 1-3; final score; all scores).

Table 4 shows the autocorrelations for lags of one to ten for each group, within each case.

Table 1. Summary of Generated Data

	Case 1		Case 2		Case 3	
	I	II	I	II	I	II
Means:						
T_1	10.04	7.04	10.04	7.04	10.04	7.04
T_{20}	26.29	31.29	26.95	31.95	26.37	31.37
all trials	22.40	22.85	22.92	23.37	22.84	23.24
Standard Deviation:						
T_1	3.16	3.16	3.16	3.16	3.16	3.16
T_{20}	3.16	3.16	3.16	3.16	3.16	3.16
Correlation	.20	.20	.80	.80	.90 → .10	.90 → .10

TABLE 2: A Comparison of Statistical Results using Difference Scores and ANOVA

STATISTICAL ANALYSES			COVARIANCE CONDITIONS AND RESULTANT STATISTICS					
Dependent Variable	Statistical Test,	Effects of Interest	Case 1 (r=.2)		Case 2 (r=.8)		Case 3 (r=.9 → .1)	
			MS error	F	MS error	F	MS error	F
D (=$T_{20} - T_1$)	2 t tests:	t^2 (group I)	0.55	497.3	0.14	2146.5	0.62	445.1
		t^2 (group II)	0.55	1107.5	0.14	4649.8	0.62	987.8
	One-way ANOVA:	G (1)*	15.9	60.4	4.0	240.0	17.9	53.5
D (=$\overline{T}_{18-20} - \overline{T}_{1-3}$)	One-way ANOVA:	G (1)	6.9	130.4	2.7	332.3	18.1	49.5
T_{20}	One-way ANCOVA: (T_1 as covariate)	G (1)	9.7	39.2	3.7	181.5	10.1	33.9
$\overline{T}_{18-20}$	One-way ANCOVA: ($\overline{T}_{1-3}$ as covariate)	G (1)	3.9	101.5	1.3	541.9	9.3	37.5
T_1, T_{20}	2 x 2 Anova:	G (1)	12.0	2.5	18.0	1.7	11.0	2.7
		T (1)	8.0	1544.6	2.0	6554.2	9.0	1381.0
		GT (1)	8.0	60.4	2.0	240.0	9.0	53.5
$T_1 \rightarrow T_{20}$	2 x 20 ANOVA with Trend Analysis:	G (1)	48.0	1.3	161.8	0.38	124.2	0.38
		T (19)	8.0	325.5	2.0	1337.5	4.0	631.0
		GT (19)	8.0	18.2	2.0	73.0	4.0	38.2
		$T_{Lin.}$ (1)	8.0	5645.5	2.0	22425.1	40.0	1032.5
		GT_L (1)	8.0	326.2	2.0	1296.8	40.0	67.1
		$T_{Quad.}$ (1)	8.0	638.7	2.0	2739.5	10.7	582.5
		GT_Q (1)	8.0	6.8	2.0	27.6	10.7	5.6

* The number in parentheses refers to the d.f. for the effect

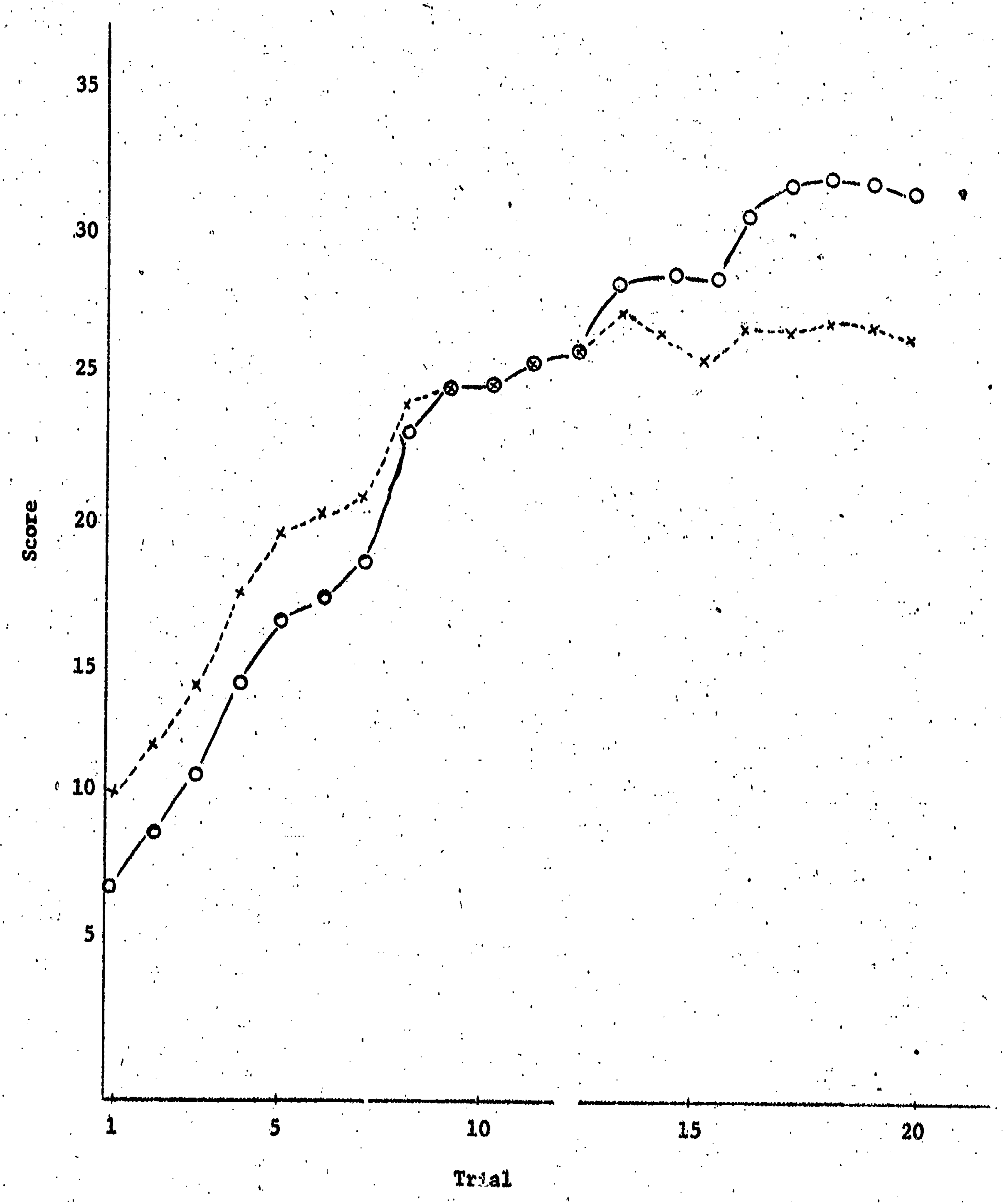
TABLE 3: A Comparison of Statistical Results using ANOVA and MANOVA

STATISTICAL ANALYSES			COVARIANCE CONDITIONS AND RESULTANT STATISTICS					
Dependent Variable	Statistical Test	Effects of Interest	Case 1 (r=.2)		Case 2 (r=.8)		Case 3 (r=.9 → .1)	
			MS error	F	MS error	F	MS error	F
$T_1 \rightarrow T_{20}$	2 x 20 ANOVA	G (1)	48.0	1.3	161.8	0.38	124.2	0.38
	with Trend	T (19)	8.0	325.5	2.0	1337.5	4.0	631.0
	Analysis:	GT (19)	8.0	18.2	2.0	73.0	4.0	38.2
		$T_{Lin.}$ (1)	8.0	5645.5	2.0	22425.1	40.0	1032.5
		GT_L (1)	8.0	326.2	2.0	1296.8	40.0	67.1
		$T_{Quad.}$ (1)	8.0	638.7	2.0	2739.5	10.7	582.5
		GT_Q (1)	8.0	6.8	2.0	27.6	10.7	5.6
$T_2 \rightarrow T_{20}$	2 x 19 ANOVA	G (1)	46.0	2.47	153.9	0.74	121.1	0.79
	with Trend	T (19)	8.0	256.6	2.0	1037.4	3.8	500.6
	Analysis:	GT (19)	8.0	17.9	2.0	71.8	3.8	39.3
		T_L (1)	8.0	4051.8	2.0	16482.3	36.5	815.0
		GT_L (1)	8.0	306.8	2.0	1221.9	36.5	69.7
		T_Q (1)	8.0	481.6	2.0	1993.3	9.4	484.5
		GT_Q (1)	8.0	3.3	2.0	13.2	9.4	3.0
$T_2 \rightarrow T_{20}$	2 x 19 ANCOVA	G (1)	39.0	9.0	33.0	64.5	82.5	10.7
	with Trend Analysis: (T_1 as covariate)	Trials	- all F values and MS errors for effects involving trials are identical to those obtained in the ANOVA example above					
$T_1 \rightarrow T_{20}$	One-way Manova:	G (20)		11.7		46.3		4.5
	- univariate on linear		8.0	326.0	2.0	1293.3	40.0	67.1
	- univariate on quadratic		8.0	6.8	2.0	27.6	10.7	5.6
	- step-down F on quadratic			1.0		1.2		3.6

Table 4. Autocorrelations on Time Series Data

Condition and Group		Autocorrelations for Lags 1 to 10									
		Lag 1	2	3	4	5	6	7	8	9	10
Case 1:	A	.75	.70	.65	.59	.54	.44	.33	.31	.26	.18
	B	.87	.85	.83	.81	.78	.73	.67	.66	.64	.57
Case 2:	A	.93	.91	.88	.84	.79	.73	.65	.60	.55	.49
	B	.97	.96	.95	.94	.93	.90	.88	.86	.84	.80
Case 3:	A	.96	.92	.86	.78	.69	.62	.50	.40	.30	.22
	B	.98	.97	.95	.92	.89	.85	.80	.74	.69	.63

Fig. 2. Graph of Performance Change



REFERENCES

Alexander, H.W. The estimation of reliability when several trials are available. Psychometrika, 12, 79-99, 1947.

Anderson, T.W., and Goodman, L.A. Statistical inference about Markov chains. Annals of Mathematical Statistics, 28, 89-110, 1957.

Bailey, N.J. The elements of stochastic processes with application to the natural sciences. New York: John Wiley & Sons, 1964.

Baltes, P.B., and Nesselroade, J.R. The developmental analysis of individual differences on multiple measures. In Nesselroade, J.R., and Reese, H.W. (eds.), Life-Span Developmental Psychology. New York: Academic Press, 219-251, 1973.

Baumgartner, T.A. Criterion score for multiple trial measures. Research Quarterly, 45, 193-198, 1974.

Baumgartner, T.A., and Jackson, A.S. Measurement schedules for tests of motor performance. Research Quarterly, 41, 10-14, 1970.

Bentler, P.M. Assessment of developmental factor change at the individual and group level. In Nesselroade, J.R., and Reese, H.W. (eds.), Life-Span Developmental Psychology. New York: Academic Press, 145-174, 1973.

Bereiter, C. Some persisting dilemmas in the measurement of change. In Harris, C.W. (ed.), Problems in Measuring Change. Madison: University of Wisconsin Press, 3-20, 1963.

Billewicz, W.Z. The efficiency of matched samples: An empirical investigation. Biometrics, 623-644, 1965.

Bock, D. Multivariate analysis of variance of repeated measurements. In Harris, C.W. (ed.), Problems in Measuring Change. Madison: University of Wisconsin Press, 85-103, 1963.

Box, G.E.P. Some theorems on quadratic forms applied in the study of analysis of variance problems: II. Effects of inequality of variance and correlation between errors in the two way classification. Annals of Mathematical Statistics, 25, 484-498, 1954.

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

Burt, C. Test reliability estimated by analysis of variance. British Journal of Statistical Psychology, 8, 103-118, 1955.

Corballis, M.C. Longitudinal factor analysis. Psychometrika, 35, 79-97, 1970.

Carlson, T. The appropriateness of the analysis of covariance to the simple-randomized design in physical education research. Presented at the National AAHPER Convention, St. Louis, Missouri, 1968.

Carron, A.V., and Marteniuk, R.G. An examination of the selection of criterion scores for the study of motor learning and retention. Journal of Motor Behavior, 2, 239-244, 1970.

Cattell, R.B. The structuring of change by P-technique and incremental R-technique. In Harris, C.W. (ed.), Problems in Measuring Change. Madison: University of Wiscousin Press, 167-198, 1963.

Cole, J.W.L., and Grizzle, J.E. Applications of multivariate analysis of variance to repeated measurements experiments. Biometrics, 810-828, 1966.

Cronbach, L.J., and Furby, L. How we should measure "change" or should we? Psychological Bulletin, 74, 68-80, 1970.

Davidson, M.L. Univariate versus multivariate tests in repeated measures experiments. Psychological Bulletin, 77, 446-452, 1972.

Dotson, C.O. Analysis of change. In Wilmore, J.H. (ed.), Exercise and Sport Sciences Reviews, Vol. 1. New York: Academic Press, 393-419, 1973.

Feldt, L.S. A comparison of the precision of three experimental designs employing concomitant variable. Psychometrika, 23, 335-353, 1958.

Feldt, L.S., and McKee, M.E. Estimation of the reliability of skill tests. Research Quarterly, 29, 279-293, 1957.

Finn, J.D. Multivariate analysis of repeated measures data. Multivariate Behavioral Research, 391-413, 1969.

Finney, D.J. Stratification, balance, and covariance. Biometrics, 373-386, 1957.

Gaito, J. Repeated measurements designs and tests of null hypotheses. Educational and Psychological Measurement, 33, 69-75, 1973.

Gaito, J., and Wiley, D.E. Univariate analysis of variance procedures in the measurement of change. In Harris, C.W. (ed.), Problems in Measuring Change. Madison. University of Wisconsin Press, 60-84, 1963.

Glass, G.V., and Maguire, T.O. Analysis of time-series quasi experiments. Project No. 6-8329, U.S. Department of Health, Education, and Welfare, Office of Education, Bureau of Research, 1968.

Gottman, J.M., McFall, R.M., and Barnett, J.T. Design and analysis of research using time series. Psychological Bulletin, 72, 299-306, 1969.

Greenhouse, S.W., and Geisser, S. On methods in the analysis of profile data. Psychometrika, 24, 95-111, 1959.

Greeno, J.G., and Bjork, R.A. Mathematical learning theory and the new "mental forestry". Annual Review of Psychology, 81-116, 1973.

Guppy, N., and Fraser, E.D. Occupational mobility in professional sport: A MARKOV process as a tool for measurement. Presented at the 1st Canadian Congress for the Multidisciplinary Study of Sport and Physical Activity, Montreal, 1973.

Harris, C.W. Canonical factor models for the description of change. In Harris, C.W. (ed.), Problems in Measuring Change. Madison: University of Wisconsin Press, 138-155, 1963.

Harris, C.W. (ed.). Problems in Measuring Change. Madison: University of Wisconsin Press, 1963.

Henry, F.M. "Best" versus "Average" individual scores. Research Quarterly, 38, 317-320, 1965.

Henry, F.M., and Demoor, J. Metabolic efficiency of exercise in relation to work load at constant speed. Journal of Applied Physiology, 2, 481-487, 1950.

Holtzman, W.H. Statistical models for the study of change in the single case. In Harris, C.W. (ed.), Problems in Measuring Change. Madison, University of Wisconsin Press, 199-211, 1963.

Hummel, T.J., and Sligo, J.R. Empirical comparison of univariate and multivariate analysis of variance procedures. Psychological Bulletin, 76, 49-57, 1971.

Jones, R.H., Crowell, D.H., and Kapuniai, L.E. Change detection model for serially correlated multivariate data. Biometrics, 269-280, 1970.

Karlin, S. A first course in stochastic processes. New York: Academic Press, 1966.

Krause, M.S. The theory of measurement reliability. Journal of General Psychology, 80, 267-278, 1969.

Kroll, W. Reliability theory and research decision in selection of a criterion score. Research Quarterly, 38, 412-419, 1967.

Lomnicki, Z.A. Some aspects of the statistical approach to reliability. J.R. Statist. Soc. A, 136, 395-419, 1973.

Lord, F.M. The measurement of growth. Educational and Psychological Measurement, 16, 421-437, 1956.

Lord, F.M. Elementary models for measuring change. In Harris, C.W., Problems in Measuring Change. Madison: University of Wisconsin Press, 21-38, 1963.

McCall, R.B., and Appelbaum, M.I. Bias in the analysis of repeated-measures designs: Some alternative approaches. Child Development, 44, 401-415, 1973.

McCraw, L.W., and McClenney, B.N. Reliability of fitness strength tests. Research Quarterly, 36, 289-295, 1965.

McNemar, Q. On growth measurement. *Educational and Psychological Measurement*, 18, 47-55, 1958.

Mendoza, J.L., Toothaker, L.E., and Nicewander, W.A. A Monte Carlo comparison of the univariate and multivariate methods for the groups by trials repeated measures design. *Multivariate Behavioral Research*, 165-177, 1974.

Merrill, M. The relationship of individual growth to average growth. *Human Biology*, 3, 37-70, 1931.

Ng, K.T. Applicability of classical test score models to repeated performances on the same test. *Australian Journal of Psychology*, 26, 1-8, 1974.

Nunnally, J.C. Research strategies and measurement methods for investigating human development. In Nesselroade, J.R., and Reese, H.W. (eds.), *Life-Span Developmental Psychology*. New York: Academic Press, 1973.

Poor, D.D.S. Analysis of variance for repeated measures designs: Two approaches. *Psychological Bulletin*, 80, 204-209, 1973.

Rosemier, R.A. The use of an exaggerated alpha in a test for the initial equality of groups. *Research Quarterly*, 39, 829-830, 1968.

Schmidt, R.A. The case against learning and forgetting scores. *Journal of Motor Behavior*, 4, 79-88, 1972.

Schutz, R.W. Stochastic processes: Their nature and use in the study of sport and physical activity. *Research Quarterly*, 41, 205-212, 1970a.

Schutz, R.W. A mathematical model for evaluating scoring systems with specific reference to tennis. *Research Quarterly*, 41, 552-561, 1970b.

Schutz, R.W. A theory of motor response organization and memory retrieval in CRT tasks. Doctoral Dissertation, University of Wisconsin, Madison, 1971.

Schutz, R.W., and Roy, E.A. Absolute error: The devil in disguise. *Journal of Motor Behavior*, 5, 141-153, 1973.

Shumway, R.H. Applied regression and analysis of variance for stationary time series. *Journal of American Statistical Association*, 65, 1527-1546, 1970.

Sidman, M. A note on functional relations obtained from group data. *Psychological Bulletin*, 49, 263-269, 1952.

Snee, R.D. On the analysis of response curve data. *Technometrics*, 14, 47-62, 1972.

Stelmach, G.E. Problems of measuring change: An overview of the problems. Paper presented at Measurement and Evaluation Symposium, AAHPER National Convention, Atlantic City, New Jersey, March, 1975.

Strahan, R.F. A coefficient of directional correlation for time-series analyses. Psychological Bulletin, 76, 211-214, 1971.

Thomas, J.R., and Chissom, B.S. Comparison of group x trials analysis of variance and discriminant analysis for use in group profile evaluations. Perceptual and Motor Skills, 37, 671-675, 1973.

Tucker, L.R. Implications of factor analysis of three-way matrices for measurement of change. In Harris, C.W. (ed.), Problems in Measuring Change. Madison: University of Wisconsin Press, 122-137, 1963.

Tucker, L.R., Damarin, F., and Messick, S. A base-free measure of change. Psychometrika, 31, 457-473, 1966.

Wiley, J.A., and Wiley, M.G. A note on correlated errors in repeated measurements. Sociological Methods and Research, 3, 172-188, 1974.

Winer, B.J. Statistical Principles in Experimental Design. New York: McGraw-Hill Book Company, 1971.