CHARACTERISTICS OF VARIABLE ERROR AND THEIR EFFECTS ON THE TYPE I ERROR RATE

by

MARC ELIE GESSAROLI

B.P.E., The University of British Columbia, 1978

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTERS OF PHYSICAL EDUCATION

in

THE FACULTY OF GRADUATE STUDIES

(PHYSICAL EDUCATION)

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

JUNE 1981

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of PHYSICAL EDUCATION

The University of British Columbia
2075 Wesbrook Place
Vancouver, Canada
V6T 1W5

Date AUGUST 4, 1981

ABSTRACT

A common practice in motor behavior research is to analyze Variable Error data with a repeated measures analysis of variance. The purpose of this study was to examine the degree to which blocked (VE) data satisfies the assumptions underlying a repeated measures ANOVA. Of particular interest was whether the assumption of covariance homogeneity - both within and between experimental groups - is satisfied in actual experimental data. Monte Carlo procedures were used to study the effect of varying degrees of violations of these assumptions on the Type I error rate.

The means and ranges of the correlation matrices of eight experimental data sets were studied for both raw and VE scores based upon different block sizes. In every situation where the experimental groups were comprised of feedback and no feedback conditions, the correlation matrix for the no feedback group displayed correlations of greater magnitudes and consistency relative to those of the feedback condition. The next phase involved using the underlying variance-covariance matrices for three of these data sets to simulate raw and VE data based on various block sizes. Raw data were simulated for each of four covariance heterogeneity conditions: (1) equality within and between the variance-covariance matrices; (2) inequality within the matrices but equality between the matrices; (3) equality within each variance-covariance matrix but inequality between the matrices; (4) inequality both within and between the two variance-covariance matrices.

Populations of 10,000 subjects for each of two groups, the underlying variance-covariance matrices being dependent upon the homogeneity of covariance condition being studied, were generated based on each of three actual experimental data sets. The data were blocked in various ways depending on the original number of trials in the experiment (36, 24 or 18) with VE being the dependent variable. An experiment consisted of randomly selecting 20 subjects for each of the two groups, blocking the trials based on specific block sizes and analyzing the raw and VE data by a repeated measures ANOVA. The effect of interest was the Groups by Blocks interaction. The complete process was replicated for the four covariance homogeneity conditions for each of the three data sets, resulting in a total of 22,000 simulated experiments.

Results indicated that the Type I error rate increases as the degree of heterogeneity within the variance-covariance matrices increases when raw (unblocked) data is analyzed. With VE, the effects of within-matrix heterogeneity on the Type I error rate are inconclusive. However, block size does seem to affect the probability of obtaining a significant interaction, but the nature of this relationship is not clear as there does not appear to be any consistent relationship between the size of the block and the probability of obtaining significance. For both raw and VE data there was no inflation in the number of Type I errors when the covariances within a given matrix were homogeneous, regardless of the differences between the group variance-covariance matrices.

ACKNOWLEDGEMENTS

I wish to thank the members of my thesis committee, Dr. A. John Petkau and Dr. A. Ralph Hakstian for their time taken in aiding me with the various, and often non-trivial, statistical problems associated with this study and Dr. Gordon L. Diewert whose enthusiam and interest in the study was greatly appreciated. Their many comments and suggestions were invaluable to the progress of the thesis. Many thanks are also given to Dr. D. Gordon E. Robertson whose knowledge of computer applications simplified many facets of my data analyses.

Special thanks and a large debt is owed to my thesis chairman and advisor, Dr. Robert W. Schutz, for his continual support, enthusiasm and leadership in all aspects of my graduate life at U.B.C. Due to Dr. Schutz, my work was not only educational, but also enjoyable. The extent of his contribution cannot be measured.

Finally, a special type of gratitude must go to my parents, whose constant support and understanding towards a seemingly endless process was, and is, very important to me.

TABLE OF CONTENTS

LIST OF TABLES

INTRODUCTION

Motor performance data is unique in that a subject is measured over numerous trials under relatively constant conditions. This large number of trials is needed due to the large intra-subject variability characteristic of most motor performance tasks. The reduction and statistical analysis of these data possesses problems not encountered in most fields of study (physiological and biological measures are usually highly reliable and, therefore, often need only one or two trials; social psychology test conditions can often not be repeated without changing the condition itself). Therefore, the purpose of this study is to examine selected problems associated with the analysis of the highly interdependent repeated measures frequently encountered in motor behaviour research.

Typical motor learning experiments require a subject to perform a number (p) of trials on a motor task, the nature of the investigation being to compare the subject's performance on that task to a predetermined target score. The difference in these two scores is called the subject's performance error for that trial. In most instances the subjects are divided into q groups based on variables such as teaching method, experimental condition, previous practice or some other factor, resulting in a q X p factorial experiment with repeated measures on the second factor. A technique known as "blocking" is often employed in an attempt to: (a) obtain a measure of intra-subject variability (VE) or; (b) smooth the data if the subjects' intertrial variability is large. Here, the p original trials are divided into c "blocks" with each block comprised of p/c

original trials. Any or all of three performance error scores are then calculated for each of the new blocks; Absolute Error (AE) - the mean absolute deviation from the target score over the p/c trials; Constant Error (CE) - the subject's mean algebraic error over the trials and, Variable Error (VE) - the square root of the within-subject variance over the trials. This blocking procedure reduces the design to a q X c factorial experiment with repeated measures on the second factor. Statistical analyses, usually analysis of variance, are then performed on each of these dependent variables with the result of interest being the Groups by Blocks interaction.

The use of these three error scores and the characteristics typical to most studies result in a number of possible problems. Absolute error, because it is an absolute value, probably has a non-normal distribution and, therefore, there may be problems when using ANOVA since the assumption of normality may be violated (Safrit, Spray & Diewert, 1980). This in itself may not be too serious since ANOVA is robust to non-normality if the number of subjects in each group is large and the population variances are equal (Boneau, 1960). However, until the distribution of AE scores, along with the underlying variance-covariance structure, has been determined, and their effects on the Type I error rate studied, the validity and interpretation of research using AE is questionable.

One of the assumptions of ANOVA is that the trials have equal variances and that all the covariances be equal to zero. Failure to adhere to this assumption results in a probability of falsely rejecting the null hypothesis greater than the set level

of significance (Box, 1954). However, it was later shown by Lana and Lubin (1963) that the Type I error rate is not inflated if the covariances are equal though not necessarily equal to zero. Constant error scores are assumed to have the characteristic of adjacent trials being highly correlated with the correlations decreasing as the trials become farther apart (Gaito, 1973; Lana & Lubin, 1963). Schutz and Gessaroli (1980), in a Monte Carlo study using data based on such a variance-covariance matrix, reported many more Type I errors than in similar studies in which fewer trials were incorporated (Collier, Baker, Mandeville & Hayes, 1967). This brings forth many questions, namely: (1) Are the number of trials under which a subject is tested related to the probability of making a Type I error? (2) How does the range of covariances affect the Type I error rate? (3) Are the magnitudes of the covariances important when using ANOVA in testing hypotheses with CE as the dependent variable? That is, does a range of covariances from 0.6 to 0.1 result in the same degree of Type I errors as covariances spanning 0.9 to 0.4? Another potential problem also associated with blocking is whether varying the block size differentially affects the Type I error rate. It has been shown that when using CE data the size of the block is of no consequence in the degree of inflation of Type I errors (Schutz & Gessaroli, 1980).

Variable error is unlike either CE or AE because it is a variance and, consequently, probably has a non-normal distribution (Safrit, Spray & Diewert, 1980). However, as with AE, this may not be serious depending upon the sample sizes and the structure of the variance-covariance matrix. In their Monte

Carlo study Schutz and Gessaroli (1980) found no inflation in the Type I error rate when VE scores were calculated from raw score matrices with unequal covariances. However, the data simulation procedures calculated VE scores which were uncorrelated across blocks. Schutz and Gessaroli used raw score covariances among trials which decreased in a linear fashion as the trials became farther apart. Mathematically, such a raw score covariance structure always will result in uncorrelated VE scores. If, in real data, the VE scores are correlated and these correlations are unequal, then problems arise when using ANOVA since the assumption of homogeneous covariances has been violated. The variance-covariance structure of empirical VE data must be studied before it can be said with any certainty if heterogeneous covariances in the raw data affect the Type I error rate.

A second part of the assumption of homogeneous covariances deals with the structure of the covariance matrices between experimental conditions. Not only do the covariances within each group have to be equal, but the magnitudes of the covariances in one matrix need to be equal to those of the variance-covariance matrices for the other group. This is referred to as "compound symmetry'. As of now it is not clear if raw experimental data have covariance matrices of this type. Extending this concept to AE, CE and VE data it is also not known if their underlying covariance matrices satisfy this assumption. It is obvious that the results obtained when analyzing AE, CE or VE by an ANOVA are, at best, inconclusive.

Over the last eight years there has been extensive debate

as to the validity of using these measures in the analysis and interpretation of motor performance (e.g., Laabs, 1973; Newell, 1976; Schmidt, 1975; Schutz, 1979). Schutz and Roy (1973) initiated this debate when they provided a mathematical proof that AE could be written as a composite score of CE and VE, but in different proportions depending upon the relative magnitude of the CE and VE scores. Absolute error is therefore redundant, and furthermore, it it is used, it can be properly interpreted: only when the CE and VE components are known. For this reason, the problems associated with the analysis of AE data will not be dealt with in this study. When analyzing CE data no "absolute answers" are available in dealing with all the potential problems but, in general, the most common difficulties have been adequately resolved by Schutz and Gessaroli (1980). Although potential problems may exist with the statistical analyses of all three error measures, AE, CE and VE, VE appears to be the least understood. Thus, this study will focus primarily on the analyses of raw and VE data.

Therefore, the purpose of this study is to: (a) discover the structure of the variance-covariance matrices associated with empirical raw and VE data; (b) determine the distributions of the raw and VE data; (c) study the effect of the number of trials on Type I errors when using VE as the dependent variable; (d) study the effect of the block size on the Type I error rate when VE is used as the measure of performance error; (e) study the effect of the degree of heterogeneity of the covariances on the probability of making a Type I error; (f) study the effect of heterogeneity of the covariance matrices between the various

groups in the experimental design on the Type I error rate.

METHODOLOGY

This study consisted of two phases - the first dealing with the analysis of empirical data and the second being a Monte Carlo study of VE data.

Phase 1

The distribution and general pattern of the variance-covariance matrices of raw and VE data were studied through the following steps:

1. Letters were sent to approximately 20 motor performance researchers requesting that they supply some of their actual experimental data from which VE was eventually calculated and analyzed (a copy of the covering letter is in Appendix A). The experimental data desired could have been learning or performance data but it had to satisfy two conditions: (a) each subject had to perform a minimum of twelve trials on a given task and; (b) each experimental condition (group) had at least twelve subjects.

Upon receipt of the data sets (eight were received) they were categorized by the type of experimental task (e.g., movement reproduction, reaction time), the level of task familiarity (learning or performance) and the experimental conditions involved (e.g., feedback, no feedback).

2. The next step involved studying the variance-covariance

structure of the raw empirical data. Correlation matrices for each data set were obtained via the statistical computer package MIDAS (Michigan Interactive Data Analysis System). Separate correlation and covariance matrices were calculated for every experimental condition within a data set. The structure of the correlation matrices was studied in the following ways: (a) the mean correlation coefficient in each matrix was calculated; (b) the maximum and minimum correlation coefficients in each matrix were noted and; (c) an inspection was conducted to see if there was a difference in the magnitude of the correlation coefficients of trials close together as compared to those farther apart. This was done by taking the mean of all correlations one trial apart, three trials apart, five trials apart, etc. In the cases where the number of trials in the correlation matrix numbered greater than forty, the mean of the correlations one, six, eleven, etc. trials apart were calculated.

3. It was imperative to discover the empirical distribution of the raw scores as this distribution would dictate the type of data to be simulated in Part 2. Histograms of the frequency distributions of every trial were obtained using the MIDAS statistical package. The main problem was to discover if the data were distributed as multivariate normal. As there is presently no easy method of testing for multivariate normality, an examination of the marginal distributions was done. Although the distributions of the marginals would not indicate multivariate normality, a departure from univariate normality would clearly make the assumption of multivariate normality

tenuous. For the purpose of this paper, data whose marginal distributions exhibited univariate normality were considered to be multivariately normally distributed.

4. The raw data were then reduced to VE scores (the size of the blocks dependent upon the number of original trials) using the Fortran computer program DATASNIFF (Goodman & Schutz, 1975). The data received from the researchers consisted of experiments having 12, 18, 20, 24, 30, 36 and 50 trials. These trials were blocked in the following manners (3 X 6 defines three blocks of six trials/block):

(a) Data set 1 - 50 trials: 10 X 5, 5 X 10

(b) Data set 2 - 24 trials: 8 X 3, 6 X 4, 3 X 8

(c) Data set 3 - 20 trials: 5 X 4, 4 X 5

(d) Data set 4 - 12 trials: 4 X 3, 3 X 4

(e) Data set 5 - 18 trials: 6 X 3, 3 X 6

(f) Data set 6 - 30 trials: 10 X 3, 6 X 5, 5 X 6, 3 X 10

(g) Data set 7 - 36 trials: 9 X 4, 6 X 6, 3 X 12

(h) Data set 8 - 20 trials: 5 X 4, 4 X 5

5. The structure of the VE variance-covariance matrices were studied for each group as outlined in step 2 above.

6. The empirical distribution of VE scores was also examined by the use of histograms as explained in step 3 above.

Phase 2

This part of the study dealt with the actual Monte Carlo procedures used to investigate the characteristics of VE and their effects on the Type I error rate. This phase consisted of generating data representing 500 experiments (two groups, 20 subjects/group; variable number of trials) for each of four

variance-covariance conditions, deriving VE scores for various block sizes, and examining Type I error rates for the Groups by Blocks interaction.

Preliminary analyses. There were two primary concerns before simulating the data. Firstly, as a computer program to simulate multivariately normal data was readily available, it was necessary to discover if the data were normally distributed. After studying the histograms of the marginal distributions of the raw scores and VE scores it was concluded that both sets of scores exhibited univariate normality based on their sample sizes. That is, from the shapes of the histograms any test of significance for normality of the marginals clearly would have failed to reject the null hypothesis.

The next step involved determining the procedure for simulating VE scores. It was essential to determine if the correlations among VE scores for the generated data would mirror those of the original experimental VE data. That is, it was essential to determine that the generated raw data be based on covariance matrices depicting actual experimental data and have correlations between blocks of VE scores similar to the actual correlation matrices of VE scores. To examine this, the variance-covariance matrix and vector of means for the raw data were specified to be exactly equal to those of an original data set (Data set 3) having 20 trials. Raw data for 20 subjects were generated. One hundred of these data sets, each having the same covariance matrix, were generated. Each data set had different raw scores due to a different "starting point" being used to initialize the data generation. The resultant data sets were

blocked and VE scores calculated in two ways: five blocks of four trials/block and four blocks of five trials/block. The net result was 100 four-by-four and five-by-five matrices of VE scores. To compare the correlation coefficients of Data set 3 with the correlation coefficients of the generated data, a "mean correlation matrix" was calculated. This matrix was obtained by calculating the mean (across the 100 matrices) of every correlation coefficient in the same position in the correlation matrix. The "mean correlation matrix" displayed coefficients of the same magnitude and range as the actual correlation matrix under both blocking conditions. Tests for differences between correlation coefficients in equivalent locations in the two matrices failed to produce significance at the .05 level. Based on these results it was concluded that generation of raw data (exhibiting multivariate normality) using an empirical correlation matrix produces correlations among VE scores which adequately reflect those in the original data.

Homogeneity conditions. The question remained as to which variance-covariance matrices to use for each group as the basis for the data generation. As these matrices are user specified, well-chosen matrices could simulate data which satisfied or violated the various assumptions involved in analyzing repeated measures data by an ANOVA.

The assumptions of ANOVA require both homogeneity of the covariances within a variance-covariance matrix as well as equality between the variance-covariance matrices depicting the different experimental conditions in the design. By specifying the nature of the matrix for each of two groups it was hoped

that the effect of violating none, one or both of these conditions could be determined when VE was the dependent variable. Therefore, four statistical conditions which span all possibilities of adherence or violation of the two variance-covariance assumptions were used as bases for the generation of raw data. The nature of the "within-group" and "between-group" covariances in each of the statistical conditions follow:

1. Condition 1 (equality within; equality between). The magnitude of the covariances within each group were equal and the magnitude of the covariances between each group were also equal. In order to obtain a variance-covariance matrix satisfying the assumption of symmetry yet reflecting the magnitude of the variances and covariances of the actual matrix the following procedures were employed: (a) the mean of the variances (diagonals) in the actual variance-covariance matrix was calculated. This value was used for all the variances in the new homogeneous matrix and; (b) the mean of the covariances (off-diagonal values) in the actual variance-covariance matrix was calculated and was used as the value to which all the new covariances were equal.

Homogeneous matrices of this type were calculated based on both Group 1 and Group 2 actual variance-covariance matrices. They were used as needed to test the effect of the violation of the two assumptions. Generated data for the two groups in Condition 1 resulted in the following variance-covariance matrices:

Group 1: The homogeneous matrix derived from the actual variance-covariance matrix of Group 1.

Group 2: Same matrix as Group 1.

2. Condition 2 (inequality within; equality between). The magnitude of the covariances within each group were heterogeneous and the magnitude of the covariances between each group were equal. The variance-covariance matrices used to generate such data were:

Group 1: The original variance-covariance matrix of Group 1.

Group 2: Same matrix as Group 1.

3. Condition 3 (equality within; inequality between). The magnitude of the covariances within each group were homogeneous and the magnitude of the covariances between each group were heterogeneous. The variance-covariance matrices used to generate such data were:

Group 1: The homogeneous matrix derived from the actual variance-covariance matrix of Group 1 (i.e., as used in Condition 1).

Group 2: The homogeneous matrix based on the actual variance-covariance matrix of Group 2.

4. Condition 4 (inequality within; inequality between). The magnitude of the covariances within each group were heterogeneous and the magnitude of the covariances between each group were heterogeneous. Generated data for the two groups resulted in the following variance-covariance matrices:

Group 1: The original variance-covariance matrix of Group 1.

Group 2: The original variance-covariance matrix of Group 2.

Selection Criteria. The primary concern of this study was to investigate the Type I error rate (using VE as the dependent variable) when the raw data had varying degrees of heterogeneity

within and between the groups in the design. It was also desirable to study the effects of the number of trials before blocking occurred on the subsequent number of Type I errors when VE was calculated. Therefore, three sets of actual experimental data (Data sets 2, 5 and 7) were used as the bases of the simulation procedures. Specifically, they were chosen based on the following design and data characteristics: (a) the range of the correlation coefficients; (b) the mean of the correlation coefficients; (c) differences in the correlation matrices between the two groups; (d) the number of trials in the experiment. All data were from learning experiments. The specific attributes of these three data sets are shown in Table I.

Simulation procedures. Five hundred two-way experiments were simulated for each of the three data sets. The number of Type I errors for the Group by Block interaction, when using raw scores and VE scores as the dependent variables, were analyzed via 500 ANOVAs for each of a number of different blocking conditions in each case. Specific procedures for each of these processes follow.

Raw scores for a population of 10000 observations having a variance-covariance matrix and vector of trial means exactly as specified by the user were generated for each group. The data were produced using the computer program UBC NORMAL (Halm, 1970). The net result was 10000 observations in each of two groups having raw scores for a specific number of trials. Samples of size 20 per group were subsequently drawn from this population. Thus the sampling was not based on an infinite

Table I

Characteristics of the Raw Experimental

Data Sets Received from Motor

Behaviour Researchers

Data Set	No. of trials	No. of S/Group	Raw		VE	
			mean r	range of r's	mean r	range of r's
1						
KR	50	30	.10	-.65 to .90	.78	.44 to .94
2						
KR	24	29	.00	-.62 to .67	.20	-.21 to .41
No KR	24	29	.83	-.46 to .95	.20	-.26 to .58
3						
KR	20	13	.05	-.57 to .78	.14	-.11 to .29
No KR	20	13	.50	-.32 to .85	.20	-.54 to .77
No KR	20	13	.20	-.54 to .77	.25	.05 to .53
4						
No KR	12	24	.20	-.66 to .70	.20	-.50 to .50
5						
KR	18	40	.00	-.57 to .60	.20	-.12 to .31
No KR	18	40	.45	-.25 to .86	.25	.02 to .44
6						
KR	30	40	.10	-.50 to .60	.10	-.50 to .60
7						
No KR	36	48	.30	-.22 to .50	.20	-.19 to .54
No KR	36	48	.15	-.36 to .48	.20	-.21 to .52
8						
KR	20	10	.35	-.65 to .88	.00	-.53 to .80
KR	20	10	.55	-.06 to .92	.70	.23 to .92
No KR	20	10	.30	-.67 to .91	.45	-.21 to .90

population. However, the sample-to-population ratio (20:10000) is sufficiently small to negate the need to incorporate any finite population correction factors into the analyses.

The data were blocked and the VE scores were calculated by a Fortran computer program (see Appendix B); the size of the block being dependent upon the number of original trials and as defined in step 4 of Phase 1 of the previous section. The validity of the calculations was tested by comparing the calculated VE scores with those produced by a program known to calculate accurate VE scores, DATASNIFF (Goodman & Schutz, 1975). The scores were accurate to the fourth decimal place.

Each experiment was analyzed by an analysis of variance on the data of twenty subjects in each of two groups. A computer program (Appendix B) read 40 subjects at a time (20 from Group 1 and 20 from Group 2) and calculated the Sum of Squares and Mean Square Error terms for all the effects, and the subsequent F value for the Groups by Blocks interaction. This calculated F value was compared to the critical F value at the .10, .05 and .01 levels of significance. The critical F values were obtained via the function subroutine UBC FVALUE which gives the F value of a user-specified level of significance based on user-specified degress of freedom. The results were stored by the computer where, after all 500 experiments had been analyzed, a table showing the number of significant interactions at the .10, .05 and .01 levels of significance was printed. The table also displayed the mean F value calculated in the 500 ANOVAs and the average Mean Square Error for each of the Groups, Subjects within Groups, Trials, Groups by Trials and Subjects within

Groups by Trials effects. The actual Type I error rate was compared to the nominal level of significance by the standard error of a proportion as given by $[p(1-p)/500]$. A difference of more than two standard errors of a proportion between the actual number of Type I errors committed and the nominal level of significance was considered significant. The net result was raw and VE scores being generated for 10000 observations in each of two groups for each of four sets of underlying variance-covariance matrices. This was done separately for each of the three data sets for each blocking condition resulting in a total of twenty-two thousand experiments being analyzed.

Effect of the number of trials. In order to study the effect of the number of initial trials on the Type I error rate when VE scores are eventually calculated, additional simulations were performed on Data set 7. Conditions 1 and 3 were studied using 12 and 24 trials as well as the actual 36 trials. It was possible to use only these two conditions since they both exhibited homogeneity of the covariances within a variance-covariance matrix. When using a specific heterogeneous variance-covariance matrix as the base it is difficult to obtain an equivalent heterogeneous matrix having fewer trials. The magnitude of the covariances were equal to those in the "mean covariance matrix" based on 36 trials. The first 12 and 24 trial means of the original data were used as the trial means for both groups in the 12 and 24 trial conditions, respectively.

The twelve trials had VE calculated based upon three blocks of four trials/block and four blocks of three trials/block while the 24 trials were collapsed into data sets of three, six and

eight blocks. Five hundred ANOVAs were performed on each of the blocking conditions as well as the original (unblocked) number of trials.

RESULTS AND DISCUSSION

Structure of the Correlation Matrices

Raw scores. It was of interest to study the patterns of the correlation matrices of the raw data and the subsequently blocked VE data for each experimental data set received. Of particular interest was whether the raw data exhibited decreasing magnitudes of the correlation coefficients as trials become farther apart as hypothesized by Gaito (1973), Lana and Lubin (1963) and others.

This pattern was common to only one of the eight data sets (Data Set 1) studied. This observation is further weakened by noting that this decreasing pattern occurred only for correlations among the first six of the fifty trials in total. The remaining correlations seemed to be randomly variable in their magnitudes. A more common occurrence (though clearly not the rule) was the magnitude of later trials being generally greater than that of earlier trials (Data Sets 2 and 5). These correlations, however, did not exhibit any particular pattern.

More striking is the difference in magnitudes of the correlations between groups of subjects who received feedback and those who did not. In almost every case where an experiment consisted of two groups, KR and no-KR, the correlation among the raw scores in the no-KR groups was much greater than for the

subjects which obtained feedback (see Table I). The sole exception, Data set 8, was based on only 10 subjects/group (less than the minimum criterion of 12 subjects/group) and, therefore, any conclusions based on this Data set are tenuous. This is most clearly exemplified by Data Set 2 . Here the mean correlation in the feedback group was approximately equal to zero while the no-KR group had an average correlation of about .83 (see Table I). Although differences between the two groups were not as extreme in the other data sets, differences still existed and were consistent regardless of the type of task (linear slide, etc.) performed. These differences can be logically explained. Subjects given feedback alter their motor program after each trial, resulting in relatively variable performances from trial to trial. However, these trial-to-trial fluctuations are not constant across subjects, thus resulting in very low correlations between pairs of trials. The no KR subjects, conversely, receive no information on which to change their responses. This results in a more consistent performance over the repeated measures.

Although the average correlation in the no KR groups is greater than for KR groups, the upper limits of the correlations are approximately equal (see Table I). In most cases, however, the lower bound of the correlations in the no feedback conditions is slightly greater than for feedback (Data Sets 2,3,5). It appears that one possible explanation for the generally higher correlations in the no KR groups is the greater correlations between initial trials. Again, this is expected since response strategies vary little in this group. Subjects,

during initial trials, probably perform with a larger degree of error. The receipt of feedback may drastically alter the response strategies and, therefore, large negative correlations between these trials result.

Variable error. When the raw data are blocked and VEs calculated the nature of the correlation matrices change. Table I displays the differences in the ranges and magnitudes of the correlations when using VE instead of the raw scores. It is obvious that there is no set pattern as to what happens to the correlations when the raw scores are blocked in different ways. Data set 1 shows a large increase in the magnitude of the correlations after VE is calculated while both Data set 5 (group 2) and Data set 3 (group 2) react oppositely.

The calculation of VE seems to increase the lower bound of the range of correlations when compared to the raw data. With the exception of Data set 3 (group 2), every data set analyzed displayed this fact. However, the opposite cannot be said for the upper limit of the correlations. Some data sets (2,3,4,5) indicate a decrease in magnitude of the upper limit of the correlations while others (Data sets 6 and 8) remain unchanged. In general, though, the effect of calculating VE is to decrease the degree of heterogeneity in the correlation matrix.

Study of the correlation matrices based on VE data revealed no specific pattern in the correlation coefficients. There appeared to be no difference in the strength of the correlation between adjacent trials compared to those farther apart. Data sets 3 and 4 had lower adjacent trial correlations than those a greater distance apart, while Data set 7, displayed the opposite

effect.

Different block sizes resulted in varying degrees of heterogeneity in the correlation matrices. Table I displays the ranges of the correlation matrices for those blocking conditions which have the largest degree of heterogeneity. Invariably, those correlation matrices corresponded to the experimental design having the largest number of blocks (i.e., the smallest block size). A general characteristic of VE data is that the range of the correlations increased inversely to the block size. The effect of these heterogeneous matrices on the probability of committing a Type I error is discussed in Condition 2 below.

Distribution of Raw and VE scores

Raw scores. Histograms plotting the raw scores for each trial in each data set suggested that raw data could be assumed to be normally distributed. That is, based on the relatively small sample sizes in each experimental group it was obvious that any test for normality (e.g.; Kolmogorov-Smirnov, Chi-square Goodness of Fit) would have failed to reject the initial assumption of normality. It is acknowledged that the small sample sizes of these data sets would result in relatively low power on any such distributional test. However, observations of the histograms failed to reveal any obvious departures from normality.

Variable error. Variable Error scores are variances and, therefore, one would expect that they are distributed as Chi-square with the appropriate degrees of freedom. Safrit et al., showed that the distribution of VE scores is dependent on

different effects under various experimental designs. These authors stated that a non-normal distribution may result, but they fell short of saying that the distribution was Chi-square. It is well known, however, that one method of making a Chi-square distribution more normal is by taking the square root of the raw scores. The VE score used in this study was the square root of the intra-subject variability within a block.

The histograms showed that VE was also distributed as univariately normal. This is understandable considering the size of the sample and the fact that the VE scores have been transformed by the square root function. Safrit et al., may be correct in stating that the theoretical distribution of VE is non-normal. However, larger samples and untransformed VE scores would be necessary to reflect this.

Violations of Covariance Homogeneity

Of major importance, statistically, is whether the analysis of data via ANOVA is valid when VE is the dependent variable. This question was studied under various degrees of violation of the homogeneity of covariances assumptions.

Condition 1 (equality within; equality between). In this condition the covariance assumptions are adhered to and, therefore, Type I error rates equal to the nominally set alphas are expected when analyzing the raw data. Table II shows that the actual number of Type I errors did not significantly differ from the nominal rate for any of the three alpha levels examined. This was consistent for all the data sets. Actual alphas which differed by more than two standard errors of a proportion from the nominal level of significance were

Table II

Proportion of Significant G X B Interactions

for Unblocked (Raw) data

Homogeneity Condition	Nominal α	Data set 2 24 trials	Data set 5 18 trials	Data set 7 36 trials
1 =within =between	.10	.118	.100	.112
	.05	.066	.058	.062
	.01	.014	.012	.010
2 ≠within =between	.10	.200*	.138*	.190*
	.05	.164*	.076*	.144*
	.01	.100*	.026*	.068*
3 =within ≠between	.10	.118	.102	.116
	.05	.056	.058	.066
	.01	.014	.010	.018
4 ≠within ≠between	.10	.178*	.110	.176*
	.05	.136*	.068	.118*
	.01	.076*	.018	.050*

*actual number of significant interactions which differ by more than two standard errors of a proportion from the nominal level of significance

classified as biased. The corresponding confidence intervals are: for α=.10, (.073≤.10≤.127); for α=.05, (.031≤.05≤.069); and lastly, for α=.01, (.001≤.01≤.019).

The analyses of VE scores display similar results (see Table III). Although, in several data sets, using VE as the dependent variable seem to decrease the actual number of Type I errors, the differences are not significant. The sole exception is in Data set 7 (12 trials) where the four blocks of three trials/block displays highly inflated Type I errors (see Table IV). No logical explanation for this is apparent. The effect of the number of original trials on the probability of committing a Type I error is explained in more detail in the discussion of Condition 3.

Therefore, as a general rule, it appears that the analysis of VE data calculated from raw data satisfying the covariance assumptions does not cause a greater number of false rejection of the null hypothesis than is expected.

Condition 2 (inequality within; equality between). Table II shows that the violation of the assumption of symmetry has the effect of increasing the probability of committing a Type I error when raw data is used. Data set 2 exhibited the greatest degree of inflation with the .01 level of significance having the largest percentage difference from the nominal alpha. At the .01 level of significance the actual proportion of Type I errors was as high as .10. The increases in the Type I error rates were 100% and 325% at nominal alphas of .10 and .05, respectively. It was expected that the raw data based on the most heterogeneous variance-covariance matrix would have the maximum level of

Table III

Proportion of Significant G X B Interactions

for VE data

Homogeneity Condition	Nominal α	Data set 2			Data set 5		Data set 7		
		3X8	6X4	8X3	3X6	6X3	3X12	6X6	9X4
1 =within =between	.10	.096	.078	.098	.082	.096	.110	.100	.086
	.05	.046	.038	.044	.048	.038	.058	.036	.036
	.01	.002	.006	.006	.006	.016	.014	.006	.000*
2 ≠within =between	.10	.088	.080	.122	.084	.094	.128*	.140*	.150*
	.05	.040	.034	.054	.044	.048	.066	.084*	.100*
	.01	.006	.010	.012	.002	.012	.026*	.042*	.034*
3 =within ≠between	.10	.098	.086	.096	.094	.082	.128*	.092	.076
	.05	.040	.036	.040	.042	.048	.062	.036	.034
	.01	.008	.008	.008	.012	.006	.014	.008	.002
4 ≠within ≠between	.10	.762*	.922*	.520*	.252*	.496*	1.000*	1.000*	1.000*
	.05	.628*	.850*	.400*	.154*	.356*	1.000*	1.000*	.998*
	.01	.380*	.658*	.218*	.048*	.172*	.998*	1.000*	.992*

*actual number of significant interactions which differ by more than two standard errors of a proportion from the nominal level of significance

Table IV

Effect of the Number of Trials on the

Type I Error Rate for Raw and VE data

Based on Data set 7

Homogeneity Condition	Nominal α	36 trials				24 trials				12 trials		
		Unblocked (Raw)	Blocked (VE)			Unblocked (Raw)	Blocked (VE)			Unblocked (Raw)	Blocked (VE)	
			3X12	6X6	9X4		3X8	6X4	8X3		3X4	4X3
1 =within =between	.10	.112	.110	.100	.086	.120	.096	.098	.094	.098	.092	.138*
	.05	.062	.058	.036	.036	.068	.054	.038	.044	.050	.046	.076*
	.01	.010	.014	.006	.000*	.014	.006	.008	.004	.010	.004	.024*
3 =within ≠between	.10	.116	.128*	.092	.076	.120	.108	.094	.092	.096	.082	.140*
	.05	.066	.062	.036	.034	.060	.042	.046	.040	.048	.042	.070*
	.01	.018	.014	.008	.002	.014	.010	.008	.004	.010	.004	.022*

*actual number of significant interactions which differ by more than two standard errors of a proportion from the nominal level of significance

inflation, which, in fact, did occur. The raw data which had the highest Type I error rate (Data set 2) was based upon a mean correlation of about 0.0 and limits of -.62 to .67. This matrix was more heterogeneous than those of both Data sets 5 and 7. However, when comparing Data sets 5 and 7 this line of reasoning was not valid. Data set 5 which resulted in the smallest increase in Type I errors had a mean correlation of approximately zero and a range from -.57 to .60. A greater number of significant F values were obtained from Data set 7, where the mean r equalled .30 and whose correlations lay between -.22 and .50. Conventional thinking would assume that the variance-covariance matrix underlying Data set 7 was less heterogeneous than that for Data set 5 and, therefore, greater inflation would occur using Data set 5. In fact, the opposite was true.

Several researchers (Box, 1954b; Gaito, 1973) have indicated that the degree of inflation increases as the number of repeated measures becomes larger. In comparing the number of Type I errors committed when using raw scores as the dependent variable for the different Data sets, it is clear that this did not always occur (Table II). However, the degree of heterogeneity of the variance-covariance matrices were not equal either. It does seem very possible that the number of Type I errors is related to the interaction of the number of trials and the heterogeneity of the underlying matrix. For example, Data set 2, which consisted of 24 trials, yielded a greater number of Type I errors than did Data set 7 which had 36 trials. However, Table I indicates that the degree of heterogeneity of the

covariances was greater in Data set 2 than in Data set 7. Therefore, it appears that the increase in covariance heterogeneity in Data set 2 more than compensates for the fewer number of repeated measures in the design and thus, more Type I errors were found with Data set 2.

The number of Type I errors found for the 36 trials of the raw scores in Data set 7 is slightly higher than in the study by Schutz and Gessaroli (1980) which employed an equal number of trials. Using a correlation matrix ranging from .54 to .95 , they found an empirical Type I error rate of about .16, .12 and .05 for the .10, .05 and .01 levels of significance, respectively, as compared to these results of .190, .144 and .068 for the same nominal alphas. The increase is probably due to the greater heterogeneity in the correlation matrix used as the basis for the generation of raw data in this study.

Analysis of the VE scores showed no significant inflation in Type I errors for Data Sets 2 and 5, but did display an inflated number of Type I errors for Data Set 7. While it is obvious that the empirical Type I error rate for all blocking conditions is well within two standard errors of a proportion for Data sets 2 and 5, certain blocking conditions in Data set 7 display actual α's outside this range. All blocking conditions displayed an increase in the number of significant Groups by Blocks interactions with the degree of inflation being greatest for the nine blocks case. The sole exception was the .05 level of significance for the 3 X 12 case where the actual Type I error rate did not differ from the nominal rate by more than two standard errors of a proportion. Table V shows the pattern of

Table V

The Mean and Ranges of the

Correlation Coefficients

for Various Block Sizes

Data Set	Blocking Pattern	mean r	range of r's
	3 X 8		
	Group 1	.35	.27 to .45
	Group 2	.28	.14 to .40
	6 X 4		
2	Group 1	.15	-.14 to .41
	Group 2	.20	-.16 to .44
	8 X 3		
	Group 1	.18	-.21 to .41
	Group 2	.18	-.26 to .58
	3 X 6		
	Group 1	.27	.19 to .36
	Group 2	.30	.24 to .40
5			
	6 X 3		
	Group 1	.12	-.12 to .31
	Group 2	.20	-.02 to .44
	3 X 12		
	Group 1	.26	.17 to .40
	Group 2	.26	.15 to .41
	6 X 6		
7	Group 1	.20	-.11 to .44
	Group 2	.17	-.06 to .51
	9 X 4		
	Group 1	.20	-.10 to .54
	Group 2	.20	-.21 to .52

the correlation matrices of the VE scores for the three block sizes of Data set 7. VE simulated with 12 trials/block had the smallest range of correlations while the VE based on four trials/block displayed the greatest heterogeneity in the correlation matrix. These results agree with previous research (Rogan, Keselman & Mendoza, 1979) in that the Type I eror rate increases as the degree of heterogeneity within a matrix increases. However, Table V shows similar degrees of heterogeneity for Data set 2, yet no inflation in the number of Type I errors occurs. Also, as the degree of heterogeneity increases as the size of the block decreases, a corresponding inflation in the Type I error rate does not occur. No viable rationale is apparent to explain these conflicting results obtained for the different data sets.

Condition 3 (equality within; inequality between).

Similar to Condition 1, the covariances within each matrix are equal, however they differ in their magnitudes between the two groups.

It seems that the assumption of equality between the covariance matrices of the different experimental conditions is quite robust if the second assumption of homogeneity within the covariance matrices is satisfied. With one exception, the empirical Type I error rate did not exceed the nominal value for any of the data sets. This held regardless whether raw scores or VE was the dependent variable.

The differences in the magnitudes of the correlations between the two groups also had no effect on the Type I error rate. Small (.15 vs .30, Data Set 7), moderate(0 vs .45, Data

Set 5) and large (0 vs .83, Data Set 2) differences in the mean correlations between groups were used with the same net result in each case - no bias in the empirical Type I error rate.

The fact that the covariances within the matrices were homogeneous allowed for an attempt to isolate the effect of the number of repeated measures and subsequent block size on the probability of falsely rejecting the null hypothesis. This was done by simulating raw data for designs having either 36, 24 or 12 repeated measures where the underlying variance-covariance matrices were equal in each case. The variance-covariance matrices satisfied the "within-group" homogeneity assumption but failed to adhere to the "between-group" assumption. Differences in the number of trials had no significant effect under this condition (Table V). Again, it seems as if the number of repeated measures is only important when the assumption of compound symmetry is violated.

The only case where the number of Type I errors committed was greater than expected was when the 12 trials condition of Data set 7 produced VE scores based on three trials per block. Here, the percentage of Type I errors found was .140 for α=.10, .070 for α=.05 and .022 for α=.01. Calculating VE using four trials/block found the number of corresponding errors to be .082, .042 and .004 - all within two standard errors of a proportion of the nominal level of significance. No logical explanation for this is apparent.

Condition 4 (inequality within; inequality between). The actual experimental variance-covariance matrices for each group were used to simulate the data for this condition. When the raw

data was analyzed the results ranged from no inflation in the number of Type I errors (Data Set 5) to serious departures for the pre-set alpha (Data Set 2). Numerous researchers, starting with Box (1954b), have shown that the probability of making a Type I error increases when the two covariance assumptions are not met. As expected, Data set 2, having the greatest degree of heterogeneity within the matrices for the two groups as well as the largest descrepancy between the matrices, has the greatest Type I error rate. However, with Data set 5, which has moderate heterogeneity both within and between the correlation matrices, the empirical level of significance failed to increases appreciably. Data set 7, having the least degree of heterogeneity both within and between the matrices, produced the second highest empirical Type I error rate (see Table II). While the last two findings contradict previous research, it must be remembered that Data set 7 had twice the number of trials (36) as did Data set 5 (18). Therefore, it again appears that when the raw data is analyzed, the degree of heterogeneity combined with the number of repeated measurements is related to the probability of committing a Type I error.

The results of the analysis of the VE data initially appear to be overwhelming because of the number of significant interactions obtained (Table III). However, this does not necessarily imply that a number of Type I errors were committed, but may reflect the fact that the VE scores between the two groups are, in fact, different. This is quite possible since subjects receiving feedback supposedly have different underlying processes on which to base their responses than do subjects who

receive no information regarding their previous response. For Data set 7, which produced almost 100% significant interactions, the actual experimental data was blocked in the same way as in the simulation procedures. Analyses of variance conducted on these original VE scores show that the two groups did in fact change differently over the blocks of trials. The calculated F for the Groups by Blocks interactions for the three blocks was 16.67, 10.51 for the six blocks and 6.60 for nine blocks. Clearly, these are all significant values. The Monte Carlo procedures produced corresponding mean F values of 20.90, 13.01 and 7.50. Although the simulated data resulted in higher F values it is quite conceivable that the actual experimental data are samples from the population on which the simulated data are based.

Effect of Block Size

The rationale for the choice of the size of the block in calculating VE scores is commonly based on practical considerations, not statistical ones. The results of this study indicate, however, that the choice of the block size may be a factor in the subsequent statistical analysis.

The most lucid example of this is for the 12 trials of Data set 7 based on the variance-covariance matrices for Conditions 1 and 3. In Condition 1 the probability of committing a Type I error differed significantly depending upon the block size chosen. At the .10 level of significance, 9.2% of the experiments had significant interactions when VE was based on four trials/block, but jumped to 13.8% when three trials/block

were used. A nominal alpha equal to .05 displayed an increase from 4.6% to 7.6% while a six-fold increase occurred (.40% to 2.4%) at the .01 level of significance. Similar changes in the number of Type I errors were found under Condition 3 for this data.

More interesting are the results of the simulations based on the actual variance-covariance matrices for each group (Condition 4). This, of course, is the one which an actual researcher would analyze. Data sets 2 and 5 both show noticeable differences in the number of Type I errors depending upon the block size used to calculate VE. In Data set 5, condition 4, the three trials/block pattern resulted in almost double the number of Type I errors found for six trials/block. The corresponding probabilities are .496 vs .252 for α=.10, .356 vs .154 for α=.05 and .172 vs .048 at the .01 level of significance. In looking at Data Set 2 (Table III) it is obvious that using four trials/block instead of three trials/block results in almost twice the number of significant interactions at the .10 level of significance and more than three times at the .01 level.

The question remains as to the nature of the relationship between the size of the block and the probability of committing a Type I error. The number of Type I errors increase inversely to the size of the block for Data set 7 (12 trials) under Conditions 1 and 3 and for Data set 7 under Condition 4. It appeared that this also was true for Data set 2 (Condition 4) since the percentage of significant interactions increased from .762 to .922 (at α=.10) as the size of the block decreased from eight trials/block to four trials/block. However, when the block

size was further reduced to three trials/block the Type I error rate decreased to .520. Therefore, the obtained results are inconclusive as to whether there is a direct relationship between block size and the probability of obtaining a signficant interaction when analyzing VE data with an analysis of variance.

Although block size is not directly related to the probability of obtaining significance for VE data, it appears that the proper choice of the size of the block may drastically affect the researcher's probability of rejecting the null hypothesis. Examining the number of significant Groups by Blocks interactions for Data set 5 under condition 4, it is apparent that the probability of obtaining significance was greater when three trials/block were used in calculating VE (Table III). In fact, at the .01 level of significance, the 6 X 3 case produced 3.6 times as many significant interactions as did the 3 X 6 blocking pattern. While the percent difference in the number of significant interactions decreases as the level of significance increases, at α=.10, the six blocks case resulted in 1.97 times the number of significant interactions as when three blocks of VE were analyzed. Similar, though not as extreme, values are apparent for the results of Data set 2, condition 4 (Table III).

The fact that the size of the blocks used to calculate VE may differentially affect the probability of achieving a significant interaction undermines the reliability of VE when it is analyzed by an ANOVA.

CONCLUSIONS

Based on the analysis of the structure of correlation matrices for raw and VE data of eight actual experimental data sets, and on Monte Carlo analyses of three of these experiments, the following conclusions can be made:

1. A "typical" correlation pattern does not exist for either the raw data or the VE scores.
2. Correlations between raw scores for subjects receiving no feedback are generally less variable and greater in magnitude than for those subjects who were given feedback.
3. The correlation matrix among VE scores is usually more homogeneous than for unblocked data.
4. Empirical performance error scores are marginally normally distributed. VE scores (the square root of the within-subject variance) also appear to have normal distributions. However, these results are based on small sample sizes (max=48) and, therefore, studies with larger samples are needed to confirm this.
5. Most empirical data sets violate both the within and between matrix homogeneity assumptions.
6. If the raw data satisfies the covariance homogeneity assumptions, then the subsequent analyses of VE scores by an analysis of variance does not inflate the Type I error rate.
7. In analyzing experiments with repeated measurements by an analysis of variance the within-group homogeneity of covariance assumption is more important than the between-group assumption. Violation of the former assumption results in an increase in the Type I error rate when raw data is analyzed but results are

inconclusive with VE data. However, when the within-group assumption is satisfied and the between-group assumption is violated no inflation in the number of Type I errors occurs.

8. The size of the block used to calculate VE affects the probability of achieving significance. Such a finding questions the reliability of using VE as a dependent measure in an ANOVA.

9. When analyzing raw data the number of trials in the design does not differentially affect the Type I error rate if the within-group correlation matrices are homogeneous. If these matrices are heterogeneous the degree of inflation of Type I errors appears to be related to an interactive effect between the number of trials and the degree of heterogeneity within the matrices.

REFERENCES

Boneau, C. A. The effects of violations of assumptions underlying the t test. Psychological Bulletin, 1960, 57, 49-64.

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 of correlation between errors in the two-way classification. Annals of Mathematical Statistics, 1954b, 25, 484-498.

Collier, R. O., Baker, F. B., Mandeville, G. K., & Hayes, T. F. Estimates of test size for several test procedures based on conventional variance ratios in the repeated measures design. Psychometrica, 1967, 32, 339-353.

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

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

Goodman, D. & Schutz, R. W. DATASNIFF: A program to check data and compute summary descriptive statistics and correlations. Developed at the Physical Education Quantification Laboratory, University of British Columbia, 1975.

Halm, J. UBC NORMAL. University of British Columbia Computing Center, 1970.

Laabs, G. J. Retention characteristics of different reproduction cues in motor short-term memory. Journal of Experimental Psychology, 1973, 100, 168-177.

Lana, R. E., & Lubin, A. The effect of correlation on the repeated measures design. Educational and Psychological Measurement, 1963, 23, 729-739.

Newell, K. M. More on absolute error, etc. Journal of Motor Behavior, 1976, 8, 139-142.

Rogan, J. C. , Keselman, H. J. & Mendoza, J. L. Analysis of

repeated measurements. British Journal of Mathematical and Statistical Psychology, 1979, 32, 269-286.

Safrit, M. J., Spray, A. J. & Diewert, G. L. Methodological issues in short-term motor memory research. Journal of Motor Behavior, 1980, 12, 13-28.

Schmidt, R. A. A schema theory of discrete motor learning. Psychological Review, 1975, 82, 225-260.

Schutz, R. W. Absolute, Constant, and Variable Error: Problems and solutions. In D. Mood (Ed.), Proceedings of the Colorado Measurement Symposium. University of Colorado Press: Boulder, Colorado, 1979, 82-100.

Schutz, R. W. & Gessaroli, M. E. The effects of block size and heterogeneity of covariance on Type I error rates with Constant error and Variable error data. in Psychology of Motor Behavior and Sport - 1979. Champaign, Illinois: Human Kinetics, 1980, 633-642.

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

Appendix A

LETTER REQUESTING EXPERIMENTAL DATA

Appendix A

LETTER REQUESTING EXPERIMENTAL DATA

24 June, 1980

Dear

As a follow-up to our Trois-Rivieres paper on heterogeneity of covariance and block size, Marc Gessaroli and I are embarking on a research project on VE. Very briefly, our research purposes are as follows: 1) determine the distribution of VE (as a variance it is probably distributed as chi-square) and ascertain how this affects the distribution of F in a typical repeated measures ANOVA; 2) examine this effect under different conditions of number of trials, blocking parameters, and variance-covariance structures. To accomplish this we plan on collecting empirical data from researchers in the field in order to determine the actual distribution of VE under various experimental conditions and for a variety of dependent variables. Based on the findings, Monte Carlo procedures will be followed to simulate reality while varying the parameters of number of trials, block size and variance-covariance structure.

As you have probably summized by now, I would like to get some of you data! We are primarily interested in learning data, but may (if not enough learning data is available) also look at performance data. What we would like is raw data (data sheets, computer listing, cards, or whatever) which has been used to reflect performance error, i.e., from linear positioning tasks, temporal accuracy, etc. We are restricting our empirical samples to data sets which meet the following requirements: 1) at least 12 trials per experimental condition, and 2) at lease 12 subjects per group (one or more groups). If you have such data set(s) available I would be most appreciative if you would send it to us. A description of the experimental design and data format, an indication of what (if any) blocking was performed, and, if possible, a copy of any published or unpublished reports of the studies would be necessary in order for us to interpret and analyze your data.

Please note - we are not conducting a review of the appropriateness of statistical analyses done in our field, and will not be re-analyzing your data (but just looking at the distribution of the raw data and the VE scores).

Marc will be using these data sets in his Master's thesis. Included on his committee are Dr. Ralph Hakstian, a noted psychometrician, Dr. John Petkau, a brilliant young mathematical statistician, and Dr. Gordon Diewert from Simon Fraser University. They all view this study as a challenging and worthy study. I believe that with their help we can make a valuable contribution to an important measurement/statistical problem in motor behavior research. Your assistance will enable us to accomplish this. We will be glad to send you a copy of our findings, and reimburse you for any costs associated with sending and duplicating materials.

Thank-you in anticipation.

Yours sincerely,

R.W. Schutz
Professor

Appendix B

PROGRAM TO CALCULATE VE AND ANOVA

Appendix B

PROGRAM TO CALCULATE VE AND ANOVA

```
      DOUBLE PRECISION X(200,18),SUBSUM(200),SUBSM2(200),TR1(100)
      DOUBLE PRECISION X2(200,100),TX2(200),BE(200),BE2(200),AE(200)
      DOUBLE PRECISION DSWGT(500),DTRIAL(500),DGRPS(500),DSWG(500)
      DOUBLE PRECISION TRT2(100), VE(200,60),TR2(100),DGXT(500),AE2(200)
      DOUBLE PRECISION TRT(100),F(500),SX,SX2,V(200,60)
      DOUBLE PRECISION FTOT,TDGXT,TDSWGT,TDTR,TDGRPS,TDSWG,SUBJ
      DOUBLE PRECISION TOTAL2,BE2G1,BE2G2,SSTR,XGT,XGTB,XTOTAL
      DOUBLE PRECISION SUBJCT,XTOT2,TOTAL,BETW,TRIALS,SUBTR,XGROUP
      DOUBLE PRECISION XGRTR,SWGT,SWG
       READ(8,16) NT,NSG,NTB,NREP,IB,FVAL10,FVAL5,FVAL1
   16  FORMAT(4(1X,I3),1X,I1,3(1X,F5.3))
        NS=NSG*2
        NTOT=NS*NREP
       NB=NT/NTB
       K=1
       L=NSG
       SX=0.
       SX2=0.
      XNB=NB
      XNSG=NSG
      XNT=NT
      XNS=NS
      L2=L/2
      K2=K+NSG
      FTOT=0.
      TDGXT=0.
      TDSWGT=0.
      TDTR=0.
      TDGRPS=0.
      TDSWG=0.
      IT10=0.
      IT5=0
      IT1=0
       DO 105 NR=1,NREP
      K=1
      L=NSG
       READ(4,1) ((X(I,J),J=1,NT),I=K,L)
       K=K+NSG
       L=L+NSG
       READ(5,1) ((X(I,J),J=1,NT),I=K,L)
    1  FORMAT(12(1X,F10.5)/12(1X,F10.5))
      IF(IB.EQ.1) GO TO 71
      NT=NB
      GO TO 106
   71  DO 11 I=1,NS
       M=0
       DO 10 J=1,NT,NTB
       J2=J+(NTB-1)
       DO 9 K=J,J2
       SX=SX+X(I,K)
       SX2=SX2+(X(I,K)**2)
    9  CONTINUE
```

```
      M=M+1
      VE(I,M)=((SX2-(SX**2)/NTB)/NTB)**.5
      SX=0.
      SX2=0.
 10   CONTINUE
 11   CONTINUE
      GO TO 201
106   DO 199 I=1,NS
      DO 198 J=1,NT
      VE(I,J)=X(I,J)
198   CONTINUE
199   CONTINUE
201   XTOTAL=0.
      K=1
      L=NS
      SUBJ=0.
      TOTAL2=0.
      BE2G1=0.
      BE2G2=0.
      SSTR=0.
      XGT=0.
      XGTB=0.
      K2=K+NSG
      L2=L/2
      DO 99 I=K,L
      SUBSUM(I)=0.
      TX2(I)=0.
      DO 98 M=1,NB
      XTOTAL=XTOTAL+VE(I,M)
      SUBSUM(I)=SUBSUM(I)+VE(I,M)
      TX2(I)=TX2(I)+(VE(I,M)**2)
 98   CONTINUE
      SUBJ=SUBJ+(SUBSUM(I)**2)
      TOTAL2=TOTAL2+TX2(I)
 99   CONTINUE
      XTOT2=(XTOTAL**2)/(XNB*XNS)
      SUBJCT=(SUBJ/XNB)-XTOT2
      TOTAL=TOTAL2-XTOT2
      DO 89 M=1,NB
      TR1(M)=0.
      BE(M)=0.
      DO 88 I=K,L2
      TR1(M)=TR1(M)+VE(I,M)
 88   CONTINUE
      BE2G1=BE2G1+(TR1(M)**2)
 89   CONTINUE
      DO 79 M=1,NB
      AE(M)=0.
      TR2(M)=0.
      DO 78 I=K2,L
      TR2(M)=TR2(M)+VE(I,M)
 78   CONTINUE
      SSTR=SSTR+((TR1(M)+TR2(M))**2)
      BE2G2=BE2G2+(TR2(M)**2)
 79   CONTINUE
      BETW=((BE2G2+BE2G1)/XNSG)-XTOT2
```

```
      TRIALS=(SSTR/NS)-XTOT2
      SUBTR=TOTAL-SUBJCT-TRIALS
      DO 69 I=K,L2
      DO 68 M=1,NB
      XGT=XGT+VE(I,M)
   68 CONTINUE
   69 CONTINUE
      DO 59 I=K2,L
      DO 58 M=1,NB
      XGTB=XGTB+VE(I,M)
   58 CONTINUE
   59 CONTINUE
      XGROUP=(((XGT**2)+(XGTB**2))/(XNB*XNSG))-XTOT2
      XGRTR=BETW-TRIALS-XGROUP
      SWGT=TOTAL-SUBJCT-TRIALS-XGRTR
      SWG=SUBJCT-XGROUP
      DGXT(NR)=XGRTR/(XNB-1)
      DSWGT(NR)=SWGT/((2*(XNSG-1))*(XNB-1))
      DTRIAL(NR)=TRIALS/(XNB-1)
      DGRPS(NR)=XGROUP
      DSWG(NR)=SWG/(2*(XNSG-1))
      F(NR)=DGXT(NR)/DSWGT(NR)
      IF(F(NR).GE.FVAL10) IT10=IT10+1
      IF(F(NR).GE.FVAL5) IT5=IT5+1
      IF(F(NR).GE.FVAL1) IT1=IT1+1
      FTOT=FTOT+F(NR)
      TDGXT=TDGXT+DGXT(NR)
      TDSWGT=TDSWGT+DSWGT(NR)
      TDTR=TDTR+DTRIAL(NR)
      TDGRPS=TDGRPS+DGRPS(NR)
      TDSWG=TDSWG+DSWG(NR)
  105 CONTINUE
      FMEAN=FTOT/NREP
      TDGXTM=TDGXT/NREP
      TSWGTM=TDSWGT/NREP
      TDTRM=TDTR/NREP
      TDGRPM=TDGRPS/NREP
      TDSWGM=TDSWG/NREP
      WRITE(6,2) IT10,IT5, IT1, FMEAN,TDGRPM,TDSWGM,TDTRM,TDGXTM,TSWGTM
    2 FORMAT('THE # OF PS LESS THAN 10 = ',I3,/,
     *'THE NUMBER OF PS LESS THAN 05 = ',I3,/,
     *'THE NUMBER OF PS LESS THAN 01 = ',I3,/,
     *'THE MEAN F VALUE WAS = ',F10.5,/,
     *'THE MEAN FOR MS GROUPS = ',F13.4,/,
     *'THE MEAN FOR SUB WITHIN GROUPS = ',F12.4,/,
     *'THE MEAN FOR MS TRIALS = ',F12.4,/,
     *'THE MEAN FOR MS GROUPS BY TRIALS = ',F10.4,/,
     *'THE MEAN FOR MS SUB WITHIN GROUPS BY TRIALS = ',F10.4)
      STOP
      END
```

Appendix C

REVIEW OF LITERATURE

Appendix C

REVIEW OF LITERATURE

Overview of Chapter

The most common experiment in motor behaviour research involves each subject performing several trials of a particular task. Repeated measures designs are invariably used since it is the researcher's goal to study how the subject performs over a period of time. In this way, some knowledge as to how a subject learns, forgets or retains may be examined. Usually, there are at least two experimental conditions in the design, thereby allowing for comparisons between various groups or treatment conditions. The data are generally analyzed by an analysis of variance.

The proper analysis of repeated measures data via ANOVA is dependent upon the data satisfying various assumptions. While the assumptions of normality and equality of variances are important and should be checked, the most common assumptions which are violated with motor learning data are those dealing with the heterogeneity of covariances. In fact, Lana and Lubin (1963) and others stated that correlations among trials closer together are larger than for those farther apart. Also, because the experimental groups are generally quite different, the covariance matrices between the various groups are probably unequal - a violation of an assumption of ANOVA. Therefore, although previous research into the effects of violating the

assumptions of normality and equality of variances will be summarized, the emphasis will be on reviewing literature concerned with the assumption of compound symmetry (homogeneity of the covariances within each group and between the variance-covariance matrices of each group). The effects of violating these assumptions on the Type I error rate and methods for compensating for covariance heterogeneity will be the main focus of this literature review.

There has been much debate in the literature over the last eleven years as to the proper choice of a dependent variable in motor behaviour studies. Some of the arguements have been made on a purely theoretical basis while others have considered the statistical properties of the dependent measures. As this study concerns itself with the analysis of one of these dependent variables (VE) a review of the ensuing debate seems appropriate.

The Statistical Model

As mentioned previously, the common motor learning experiment consists of each subject performing several trials (q) of a specific task. Usually the subjects are divided into p experimental groups, the resultant design being a p X q experimental design with repeated measures on the last factor. This data is subsequently analyzed by an analysis of variance.

The model underlying a repeated measures ANOVA of this type is linear in nature and defined by:

$$X_{ijk} = \mu + \alpha_j + \beta_k + \pi_{i(j)} + \alpha\beta_{jk} + \beta\pi_{ki(j)} + \epsilon_{ijk}$$

where X_{ijk} defines the score for the ith subject in the jth group on the kth trial; μ is the overal population mean; α_j and β_k are the effects of the jth treatment and the kth occasion,

respectively; $\pi_{i(j)}$ is a constant relating the ith subject with the jth treatment group; $\alpha\beta_{jk}$ is the interaction of the jth group with the kth occasion; $\beta\pi_{ki(j)}$ is the interaction of occasion k and subject i within j; and ϵ_{ijk} is the random error in the system. Furthermore, these parameters are subject to the following constraints:

$$\sum_j \alpha_j = \sum_k \beta_k = \sum_j \alpha\beta_{jk} = \sum_k \alpha\beta_{jk} = \sum_k \beta\pi_{ki(j)} = 0,$$

where i=1,...,N; j=1,...,P; k=1,...,Q

Assumptions of Repeated Measures ANOVA

The specific assumptions underlying the analysis of repeated measures data by analysis of variance are as follows:

1. The populations must be multivariately normally distributed.
2. The population variances must be equal.
3. (a) The magnitudes of the covariances within a group must be equal.

(b) The magnitudes of the covariances between each grouping factor must be equal.

Assumption of normality. The first assumption underlying an analysis of variance is that the populations must be distributed as multivariately normal. However, as tests for multivariate normality are few and somewhat complex in nature (see Gnanadesikan, p. 151-195), the less stringent assumption of univariate normality between the marginal distributions has been accepted as a satisfactory condition for a valid F test. Several early pieces of research have been done studying the effects of non-normality on the probability of committing a Type I error. Although a multitude of research regarding the effects of non-normality exists, only a summary of the conclusions will be

presented here.

Boneau (1960), using equal sample sizes and equal variances found that an ANOVA is quite robust to varying levels of non-normality. In fact, inflation in the number of Type I errors was found only when one or more of the populations were non-normal (i.e., exponential or rectangular) and the sample sizes were very small (five subjects/group). As the sample sizes increases to 15 subjects/group, the actual number of Type I errors was only slightly higher than the nominal value. Scheffe (1959) has proven mathematically that the robustness of the ANOVA F test increases as N becomes large with F tests being perfectly robust with infinite sample sizes. Therefore, it appears that if the sample sizes and variances are equal, the F test is quite robust to violations of the normality assumption with the robustness increasing as N increases.

When non-normality is combined with other factors such as unequal variances and/or covariances the results are different. Several investigators have stated that ANOVA is fairly robust to departures from normality and equality of variances (e.g., Gaito, 1973; Wilson & Lange, 1972), but Bradley (1980) showed that the combination of these two factors severely affects the Type I error rate. Bradley, attempting to simulate real-life data, found that under varying levels of unequal sample sizes, non-normality and variance ratios, 25% of the situations failed to produce a reasonable F level when N was less than 100. He found that the sample size needed for robustness increased as the level of significance decreases. More specific conditions and their effects on the robustness of the test are discussed in

the article.

Non-normality when combined with covariance heterogeneity has the effect of inflating the Type I error rate of the within-subjects main effect, especially when using multivariate tests (Mendoza, Toothaker & Nicewander, 1974; Rogan, Keselman & Mendoza, 1979). However, when the effect of interest was the within-subjects interaction, the actual Type I error rate underestimated the nominal level of significance. Thus, when analyzing within-subjects effects from non-normal data displaying heterogeneous covariances, the effect being tested must be considered.

Homogeneity of variances. An early study by Hsu (1938) showed that the t-test is robust to inequality of variance if the sample sizes are equal. However, the actual probability of committing a Type I error moves away from the nominal level of significance as the ratio between the variances and/or the degree of inequality between sample sizes increase (Hsu, 1938; Scheffe, 1959). More specifically, when the smaller variance is associated with the larger population, an inflation in the Type I error rate occurs while in the situation where the larger population has the larger variance, the actual alpha underestimates the nominal level. Collier, Baker, Mandeville and Hayes (1967), in a Monte Carlo study, found that there were no extreme departures from the nominal alpha levels if the covariances and sample sizes were equal and any inflation which did occur decreased as the sample size increased.

As with the assumption of normality, the F test is quite robust to violations of the homogeneous variances assumption

with the degree of robustness increasing with increases in the sample size. However, as Bradley (1980) displayed, the interactive effects of the violations of the various assumptions can have severe effects on the Type I error rate, and the fact that the sample sizes are equal is not sufficient reason to assume robustness of the ANOVA.

Homogeneity of covariances. The final two assumptions can be represented by the Q X Q population variance-covariance matrix of the form:

$$\begin{bmatrix} \sigma^2 & \rho\sigma^2 & \cdots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \cdots & \rho\sigma^2 \\ \vdots & \vdots & & \vdots \\ \rho\sigma^2 & \rho\sigma^2 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}$$

Defining the population variance-covariance matrix for each level of P as Σ_j, the above matrix must be common to all levels of P (i.e., $\Sigma_j = \Sigma$, $j=1,\ldots,P$) in the p x q design. A matrix of the above form is said to have the properties of "compound symmetry" or "uniformity" (Geisser, 1963) or "multisample sphericity" (Huynh, 1978). Studies in motor learning in which a subject is tested on many trials over time on a task, in most cases, do not adhere to the equality of covariance assumptions. It is not unlikely to have higher correlations between adjacent trials with the magnitude of the correlations decreasing as the trials become farther apart (Davidson, 1972; Greenwald, 1976; Lana & Lubin, 1963; Wilson, 1975). The question remains as to the effect on the validity of the ANOVA when one or more of the above assumptions are violated. This study is primarily concerned with the effect of the the violations of these

assumptions and, therefore, the remainder of the literature review deals almost exclusively with variance-covariance heterogeneity problems.

Heterogenity of Covariances: Definition and Measurement

A measure of covariance heterogeneity. Box (1954b) in studying the effects of unequal covariances on a one-way test for differences in treatments found that the ratio, SS_T/SS_{WITHIN}, has an approximate F distribution with degrees of freedom equal to $(q-1)\epsilon$ and $(q-1)(p-1)\epsilon$, where ϵ is defined as

$$\epsilon = q^2(\bar{\sigma}_{tt}-\sigma..)^2/(q-1)[\sum\sum\sigma^2_{ij} - 2k\sum\sigma^2_{i.} + k^2\sigma^2_{..}],$$

and $\bar{\sigma}_{tt}$ is the mean of the column variances, $\sigma_{i.}$ is the mean of the ith row and $\sigma_{..}$ is the mean of all the elements in the population covariance matrix.

Geisser and Greenhouse (1958), in extending Box's findings, showed that ϵ must lie between $1/(q-1)$ and 1. If the variances are homogeneous and the covariances are homogeneous, $\epsilon=1$. Extreme degrees of heterogeneity result in ϵ having a value of $1/(q-1)$. Under the condition of complete homogeneity amongst the variances and covariances ($\epsilon=1$) the degree of freedom for the critical F are $(q-1),(n-1)(q-1)$, while when $\epsilon=1/(q-1)$ the test statistic for significance is $F[1,(n-1)]$. As is obvious, the former F value is less stringent than the latter, therefore, it is called a "liberal" test while the latter critical F value is greater resulting in a "conservative" test.

Applying Box's results from a one-way classification to the

two-way classification (i.e., a grouping factor exists) Geisser and Greenhouse (1958) found that the adjusted degrees of freedom corresponding to the test for significance between treatments (MS_T/MS_{SwGT}) and for interactions (MS_{GxT}/MS_{SwGT}) to be $(q-1)\epsilon$, $p(n-1)(q-1)\epsilon$ and $(p-1)(q-1)\epsilon$, $p(n-1)(q-1)\epsilon$, respectively. The upper and lower bounds for ϵ in the two way classification remain at 1 and $1/(q-1)$, thus facilitating the calculation of the liberal and conservative critical F values.

Compound symmetry and circularity. In a Groups by Trials repeated mesures design three test statistics (F ratios) are calculated by the ANOVA: $F_G = MS_G/MS_{SwG}$, a test for differences in groups; $F_T = MS_T/MS_{SwGT}$, a test for differences between trials; and $F_{GT} = MS_{GxT}/MS_{SwGT}$, a test for interaction. All these ratios are distributed as F with appropriate degrees of freedom if compound symmetry exists in the variance-covariance matrix (with the exception of MS_G/MS_{SwG} which is not dependent upon such a restriction). Similarly, if there is no grouping factor (i.e., a one-way classification), the test statistic for differences in treatments is given by $F = MS_T/MS_{WITHIN}$. F has an F distribution if the compund symmetry assumption is satisfied.

Work by Rouanet and Lepine (1970) and Huynh and Feldt (1970) has shown that the assumption of uniformity or symmetry of the variance-covariance matrices need not necessarily be met for the F ratio to be legitimate. Given q trials in a one-way classification, a sufficient condition for an exact F test is when $C'\Sigma C = \sigma^2 I_{(q-1)}$, where Σ is the population variance-covariance matrix, I is the identity matrix and C is a (q-1)-dimensional orthonormal contrast matrix (Huynh & Feldt, 1970;

Rouanet and Lepine, 1970). Both sets of authors, by different methods, show that if this condition is met (the condition is defined as "circularity") the Box-Geisser-Greenhouse correction factor ϵ is equal to one. Extending this idea, it follows that if the symmetry assumption is satisfied then so is the circularity assumption. However, it does not necessarily follow that circularity implies symmetry of the variance-covariance matrix (Rouanet & Lepine, 1970). It is obvious that circularity is a less stringent requirement necessary to obtain valid F ratios by analysis of variance.

In a two-way classification (a between-groups factor exists) the condition which must be satisfied is: $C'\Sigma_p C=\sigma^2 I$, p=1,...P (Huynh & Feldt, 1970). This implies that the condition of circularity, as described above, exists for each of the P groups and that the value of $C'\Sigma C$ results in the same value of the scalar, σ^2 , for each group.

The primary difference between the results of Huynh and Feldt and those of Rouanet and Lepine is that the first set of authors deal only with the circularity conditions for the overall F test while Rouanet and Lepine consider both overall and partial F tests. Rouanet and Lepine showed that certain partial comparisons are valid even if the overall circularity condition is not satisfied. The example given by Rouanet and Lepine is based upon a four by two classification with repeated measures on both factors (i.e., eight treatments). They define the overall comparison (7 df) as well as three possible partial comparisons based upon the two factors (3 and 1 df) and the interaction (3 df).

Two methods have been suggested for testing partial comparisons: (1) using an error term based upon the corresponding sum of squares as the effect being tested. For example, the denominator for the test of a comparison based on factor A would be Subjects within A. The corresponding F ratio is designated as F'. (2) Using an error term based upon the overall sum of squares (i.e., the sum of the three sum of squares, SS_{SwA}, SS_{SwB}, SS_{SwAB}). This ratio is called F".

Authors differ in their opinion as to which is the proper error term to use. Many texts favor the us of F" only while others state that F' should be used in all cases (e.g., Gaito & Turner, 1963). Since the degrees of freedom are larger in the error term for F" than for F', it would seem that F" yields a more powerful test. However, satisfying the circularity assumption for F" is more difficult than for F'. If the overall circularity assumption is satisfied (F" is valid), then any of the partial comparisons (F') are also valid. However, the opposite does not apply. The assumption for F' is less stringent than for F" and becomes weaker as the degrees of freedom in the error term decrease. Furthermore, even the stricter condition of overall circularity is less rigorous than the classical symmetry assumption.

As the F test is not valid if circularity assumptions are not met, it is necessary to be able to test for circularity. Huynh and Feldt (1970) provide a test for overall circularity based on the Box test (1950) and Mauchly's criterion W (1940). The statistics calculated are similar to those in testing for symmetry in the variance-covariance matrices. Rouanet and Lepine

(1970) adopt a multidimensional approach in testing for circularity based upon an adaptation of Anderson's 'sphericity test' (1958, p.263). Although Rouanet and Lepine do not give a test when there are p between-level factors or groups, Box's (1950) test could be used to test $C'\Sigma_p C=\sigma^2 I$, $p=1,\ldots,P$. If the null hypothesis is not rejected, Andersons's test (1958) could subsequently be employed.

Covariance Heterogeneity and Type I Error Rates

Evidence of Type I error inflation. Several investigators (e.g., Box, 1954a,b; Collier, Baker, Mandeville & Hayes, 1967; Gaito, 1961; Geisser and Greenhouse, 1958; Lana & Lubin, 1963) have discussed the effect of covariance heterogeneity upon the Type I error rate. Kogan (1948) was the first to postulate that when the trials were positively intercorrelated the subsequent F test for differences in the trials would be liberal. Box (1954b) investigated the situation where adjacent trials had correlations equal to zero. He found that the probability of obtaining a significant p-value increased as the correlations increased from 0 to ±.40. As the magnitude of the correlation increases the value of ϵ decreases. When $r=0$, $\epsilon=1$; with little correlation ($r=.20$), $\epsilon=.9507$ and a correlation of .40 resulted in ϵ equalling .8033. The corresponding negative correlations resulted in epsilon values of .9640 and .8862. Negative correlations have less of an effect on the Type I error rate than do their positive counterparts. Box concluded that as the value of ϵ decreased the probability of falsely rejecting the null hypothesis increased. Gaito (1973) calculated epsilon

values for correlations greater than .40 for a covariance structure similar to that of Box (1954b) and found that epsilon decreased quite rapidly as the correlation increased (e.g., r=.60, ϵ=.5977; r=.80, ϵ=.4009; r=.90, ϵ=.3189). He found the Type I error rate increased similarly, with a correlation of +.90 resulting in an actual probability of making a Type I error of .16 at the .05 level of significance.

Collier, Baker, Mandeville and Hayes (1967) studied several very simple covariance matrices having high adjacent trial correlations with the magnitudes of the correlations decreasing as the trials become farther apart. Using only four trials and correlations ranging from .80 to .20, they found the p-levels to be about twice as large as the expected .05 and three to five times as large at the .01 level of significance. It is quite possible that many studies have more than four trials and the subsequent error rate could be much higher than those reported by Collier et al., (1967). Schutz and Gessaroli (1980) used a correlation matrix with a similar magnitude and pattern of correlations but had data for each of 36 trials. Their Monte Carlo study resulted in a Type I error rate of .17 at the .10 level, .12 at an alpha of .05 and .05 at the .01 level - a degree of inflation greater than that of Collier et al., (1967). This is consistent with results of Box (1954b) who discovered that the value of epsilon decreases inversely with the number of trials.

Wilson (1975), in a simulation study based on each "subject" having 10 trials, used an arbitrary correlation matrix with the correlations ranging from 0 to 0.98. The Type I error

rate was consistent with the high degree of covariance heterogeneity and moderate number of trials. At the 5% level of significance the actual Type I error rate was over 20% and at the 1% level it was about 13%.

Traditional adjustments in the degrees of freedom.

Several methods have been suggested to deal with the problems produced by covariance heterogeneity; some are methodoligical; some focus on the choice of statistical test, and others try and reduce the bias in the F ratio by altering the degrees of freedom.

Greenhouse and Geisser (1959) based on the previous work of Geisser and Greenhouse (1958) and Box (1954b) proposed a three step procedure in analyzing repeated measures experiments. They suggested first doing a conservative F test. This involves using the lower bound of epsilon, $1/(q-1)$, where q is the number of trials, thereby making the adjusted degrees of freedom 1 and (N-1) d.f. for the test of a trials effect. In the groups by trials design the conservative test for an interaction effect would be distributed as F with 1 and $p(n-1)$ degrees of freedom, where p is the number of groups and n is the number of subjects under each level of p. If this proved significant, the test would be finished. If, however, the null hypothesis was not rejected, then an F test based on the conventional degrees of freedom ($\epsilon=1$) should be done. Here the degrees of freedom corresponding to the tests for a trials effect and group by trials interaction would be $(q-1),(q-1)(n-1)$ and $(q-1),p(q-1)(n-1)$, respectively. If the F ratio is non-significant the testing is finished. If the situation arises where the conservative test proves non-

significant and the conventional test significant, then an attempt must be made to estimate ϵ. The exact value of ϵ would give the actual distribution of the F ratio.

Studies using ϵ. As is obvious from the earlier equation defining epsilon, ϵ can be calculated only if the population variance-covariance matrix is known. In actual experimental data the population values are never known. Geisser and Greenhouse calculated the sample estimate ($\hat{\epsilon}$) of ϵ in the same manner as the original equation, with the population variances and covariances being substituted by the corresponding sample statistics. The degrees of freedom of the critical F are then reduced using $\hat{\epsilon}$ rather than ϵ.

Several studies have investigated the effect of using $\hat{\epsilon}$ instead of ϵ in controlling for Type I errors. Collier et al., (1967) found that, in general, $\hat{\epsilon}$ was a good estimate of epsilon. However, $\hat{\epsilon}$ is a conservative estimate of ϵ when the population value is near one resulting in a somewhat conservative test of the null hypothesis. The sampling distribution of ϵ is negatively skewed at its upper limit but becomes less variable and less biased as the population value decreases (Collier, Baker, Mandeville & Hayes, 1967; Mendoza, Toothaker & Nicewander, 1974; Rogan, Keselman and Mendoza, 1979; Stoloff, 1970; Wilson, 1975). Stoloff (1970) reported data which indicated that, as the sample size increases, the test using $\hat{\epsilon}$ to adjust the degrees of freedom results in the empirical Type I error rate is closer to the nominal rate when ϵ is approximately one. The difference in Type I errors using ϵ and $\hat{\epsilon}$ decreases as the sample size increases and

as ϵ decreases (Collier et al., 1967; Stoloff, 1970). An interesting aspect of Stoloff's study is how ϵ and $\hat{\epsilon}$ react when the number of trials increased. He found that as the trials increased, the magnitude of the Type I errors increased when ϵ was used as the correction factor. However, when the degrees of freedom were reduced by $\hat{\epsilon}$, the probability of making Type I errors decreased. This was consistent under varying levels of ϵ. It appears that the sample estimate of epsilon controls the Type I error rate better than the population value as the number of trials increase. As the maximum number of trials used was five, further investigation should be undertaken to see how conservative the test using $\hat{\epsilon}$ becomes as the levels of the repeated factor increase to a much higher degree.

Modifications of ϵ : $\hat{\epsilon}$ and $\tilde{\epsilon}$. The fact that the value of epsilon based upon sample data is negatively biased at high levels of ϵ caused Huynh and Feldt (1976) to develop a new statistic to adjust the degrees of freedom in the F ratio. This estimator, $\tilde{\epsilon}$, eliminates most of the negative bias in the test for significance when ϵ is used. They define $\tilde{\epsilon}$ as:

$$\tilde{\epsilon} = [n(k-1)\hat{\epsilon}-2]/(k-1)[n-1-(k-1)\hat{\epsilon}]$$

for the one-way classification with k trials and, for the groups by trials design:

$$\tilde{\epsilon} = [N(k-1)\hat{\epsilon}-2]/(k-1)[N-g-(k-1)\hat{\epsilon}],$$

where N is the total number of subjects and g is the number of groups. In the latter design, $\hat{\epsilon}$ is calculated by using the pooled estimates of the sample variance-covariance matrices for each of the g groups. This, of course, assumes that all the individual population variance-covariance matrices are equal for

all the groups. Huynh (1978) deals with the case when this is not true. Huynh and Feldt (1976) note that for any values of n and k, $\tilde{\epsilon}$ is always greater than $\hat{\epsilon}$, with this difference decreasing as n increases. This formula for $\tilde{\epsilon}$ allows it to have a value greater than one when there is a high degree of homogeneity in the matrix. In this case, the upper limit is exceeded. Therefore, $\tilde{\epsilon}$ is equated to one if the actual calculation of $\tilde{\epsilon}$ is greater than one. Huynh and Feldt (1976), in a Monte Carlo study comparing $\hat{\epsilon}$ and $\tilde{\epsilon}$ in controlling for Type I errors under varying levels of ϵ ($.363 \leq \epsilon \leq 1.000$) found that, in general, $\tilde{\epsilon}$ is the better estimator when ϵ is greater than 0.75 while $\hat{\epsilon}$ is superior at higher degrees of heterogeneity. They also discovered that both tests behave differently depending upon the number of groups and subjects. They state, "It can be seen that the test based on ϵ is more satisfactory when the parameter is relatively low or when the number of blocks or subjects is fairly large. The test based on $\tilde{\epsilon}$, on the other hand, behaves very well at the nominal ten or five per cent levels in all of the situations considered. At the nominal 2.5 and 1 percent levels it gives somewhat more relaxed, but reasonably adequate, control over Type I error whenever the covariance matrix is not extremely heterogeneous. This test is less dependent on the number of blocks, and is fairly good even with a block size as small as twice the number of treatment levels." (p. 80)

GA and IGA tests. Huynh (1978) extended the work of Huynh and Feldt (1976) to consider the case when the various population matrices are heterogeneous. Two tests, the General

Approximate test (GA test) and the Improved General Approximate test (IGA test) were developed to deal with this situation. The GA and IGA also have the added flexibility of being suitable for tests with unequal sample sizes. Huynh (1978), comparing all four tests ($\hat{\epsilon}$, $\tilde{\epsilon}$, GA and IGA) in a situation where the matrices almost exhibited multisample sphericity found that the GA and $\hat{\epsilon}$ approximate tests always err on the liberal side. However, the IGA and $\tilde{\epsilon}$ tests yielded better overall control of the Type I error rate. Huynh then compared the IGA and $\tilde{\epsilon}$ tests under eleven different heterogeneity conditions with the result that the IGA test tended to function better than the approximation, although both were slightly liberal. However, most differences were at smaller levels of significance or when the sample sizes were quite large (N=30). Huynh concludes that although the IGA test is more accurate and flexible, it is computationally more complex and, in many situations, the $\tilde{\epsilon}$ approximate procedure functions as well as the IGA test and, therefore is more desirable.

Multivariate techniques. An alternative to the various correction techniques applied when repeated measures data is anlayzed by an analysis of variance is a multivariate analysis of variance (MANOVA). Multivariate analysis of variance, which requires no assumptions of within-group variance or covariance homogeneity, has been frequently recommended as the appropriate technique for all repeated measures designs (Davidson, 1972; Morrow & Frankiewicz, 1979; Schutz, 1978).

Among the basic assumptions in multivariate analysis of variance are: (a) the data are distributed as multivariate

normal, and (b) the group covariance matrices all come from a single population covariance matrix. However, while MANOVA has less stringent assumptions, violations may have serious consequences on the Type I error rate.

The effects of violating the assumption of normality are generally not severe. Mardia (1971) and Ito (1969) found that the multivariate tests are quite robust to departures from multivariate normality, especially if the sample sizes are equal. Studies investigating the assumption of equal covariance matrices between groups found that the Type I error rate is controlled under moderate degrees of heterogeneity if the sample sizes are equal (Holloway & Dunn, 1967; Hakstian, Roed & Lind, 1979; Ito & Schull, 1964; Rogan, Keselman & Mendoza, 1979). Holloway and Dunn, however, found that sample size equality does not necessarily ensure control of the number of Type I errors committed as the ratio of the sample size to the number of dependent variables and the degree of covariance heterogeneity are also important. Using a ratio of 10:1 between the variances in the two covariance matrices, Holloway and Dunn discovered that equal sample sizes of 25 were sufficient when only two or three variates were used but, for 10 variates, the multivariate test, Hotelling's T^2, was not robust until the sample reached 100. In relating these results to actual behavioral data, it must be remembered that a realistic extreme for the ratio between population variances is only 2.5 (Hakstian, Roed & Lind, 1979). Hakstian et al., (1979), using variance scale factors up to 2.5 showed that the T^2 procedure was relatively robust to violations in the covariance assumption, even when the ratio

between subjects and dependent variables was as low as 3:1. While the test of main effects appear to be relatively robust, other multivariate procedures testing for significant interactions did not show the same results.

In studying the effects of covariance heterogeneity (with equal sample size, ratio of subjects to variates approximately 4:1) on the tests for significant interactions, Rogan, Keselman and Mendoza (1979) discovered an inflation in the number of Type I errors. These increases were slight for the Pillai-Bartlett trace criterion, and Wilk's likelihood ratio criterion, but were much larger (as high as .070 at alpha equal to .05) when Roy's largest root criterion was used.

When unequal sample sizes exist, the Type I error rate fluctuates greatly, with the Type I error rates increasing quickly to very unacceptable levels as the degree of heterogeneity increases, even at small sample size ratios as low as 2:1. In the most extreme case studied, with 10 variates, 50 subjects in one group compared to 10 in the other, and the variances in one group scaled at 2.5 times the magnitude of the other group, the Type I error rates were: for α=.01, .152; for α=.05, .337 and; for α=.10, .473 (Hakstian et al., 1979). Clearly, as the authors point out, "the T^2 procedure is not robust in the face of covariance matrix heterogeneity coupled with unequal n's, even for relatively minor departures from equality of the covariance matrices, sample sizes or both." (p. 1261)

Overview of univariate vs multivariate tests on power. In general, when the univariate assumptions regarding the

covariance matrices are met, the conventional univariate ANOVA is more powerful than multivariate techniques (Mendoza, Toothaker & Nicewander, 1974; Rogan et al., 1979). Of interest is the comparison between the power of the adjusted univariate tests (e.g., $\hat{\epsilon}$, $\tilde{\epsilon}$), the conventional univariate test and multivariate tests under various levels of within-group and between-group covariance matrix heterogeneity. When all covariance assumptions are met the conventional univariate test is more powerful than either the adjusted univariate tests or the multivariate tests. However, as the degree of within-group matrix heterogeneity increases the multivariate tests become more powerful in detecting significance for differences in the main effects. Rogan et al., (1979), found that as the value of ϵ decreased the power of all the tests decreased, but the multivariate tests decreased at a slower rate. As the degree of covariance heterogeneity increases the power of the adjusted univariate tests are of concern since they are the test of significance. It appears that when epsilon dips below .75 the multivariate tests more often detect the differences in the means (Mendoza et al., 1974; Rogan et al., 1979). When $\epsilon \geq .75$ the adjusted univariate tests are more powerful than their multivariate counterparts.

Mendoza et al., (1974), found that the power of detecting small interactions was greatest for Roy's largest root criterion but in detecting large differences, the adjusted univariate tests were more powerful (for $\epsilon < .75$). Rogan et al., examined the power of three multivariate tests for interaction and reported similar results as in the test for main effects, that being that

the multivariate tests were more powerful that the univariate tests. It should be noted, however, that Roy's largest root criterion had the greatest Type I error rate under covariance heterogeneity and, caution must be employed if it is to be used.

Summarizing, if all the covariance assumptions are met, the conventional univariate test is the best to use in testing for both interactions or main effects. With moderate levels of heterogeneity in the covariance matrices ($\epsilon \geq .75$) the adjusted univariate tests are best and, generally, when $\epsilon < .75$ the multivariate tests are the most powerful.

Summary. When dealing with data which exhibits heterogeneity of covariances (as is common in repeated measures behavioral data) the easiest, and often sufficient method of correcting for this heterogeneity is to use the three-step procedure as outlined by Geisser and Greenhouse (1959). However, if a sample estimate of ϵ need be calculated to adjust the degrees of freedom there are several choices. If ϵ is less than .75 the best univariate statistic is $\hat{\epsilon}$, but if epsilon is greater than .75 either $\tilde{\epsilon}$ or IGA approximate tests are the most powerful yet control for the Type I error rate. Of the latter two, the $\tilde{\epsilon}$ is much easier to calculate and is quite often as good in controlling for Type I errors as the IGA test. Multivariate tests do not depend upon the assumption of within-group covariance homogeneity and, as such, may often be the preferred method of analysis. They prove to be more powerful than their univariate counterparts when $\epsilon < .75$ but are weaker above this level. Multivariate tests, however, do require that the covariance matrices between groups come from a common

population matrix, an assumption which may not be often satisfied in motor behavior research.

AE-CE-VE Debate

A considerable controversy has developed in the past ten years regarding which statistics (AE, CE or VE) should be used as mesures of a subject's performance on some motor performance task. The debate has been primarily between those researchers who are concerned with the statistical and mathematical properties of AE, CE and VE and those investigators who are more interested in the conceptual interpretation of these scores. An excellent review of this debate is given by Schutz (1979). While several researchers had previously commented on the appropriateness of these performance measures (Burdick, 1972; Laabs, 1973; Schmidt, 1970; Underwood, 1957; Woodworth, 1938) the problem received serious consideration after a paper by Schutz and Roy (1973) proved mathematically that AE is directly related to CE and VE and, as such, can only be interpreted in light of the latter two measures. They stated that all the information of AE is found in CE when the ratio of $CE/\sqrt{VE}$ is greater than 2.0 or is in VE when CE is approximately equal to zero. AE is a weighted combination of CE and VE when $0<CE/\sqrt{VE}<2.0$. As the mathematical derivations, discussed above, were based on the assumption that the raw performance scores are normally distributed, the validity of their conclusions decreases as the departure from normality increases.

The use of Variable error as the optimal measure of within-subject variablity was questioned by Burdick (1972) and Schutz, Roy and Goodman (1973) because it did not reflect the temporal

dimension of performance errors. An alternate choice of measures such as the Mean Square Successive Difference, the Autocorrelation and the Coefficient of Temporal Variablity have been suggested by Burdick (1972). Schutz et al., (1973) suggested that the non-normal distribution of a variance results in a loss of power when VE is analyzed by an ANOVA and indicated that the autocorrelation coefficient be used as an additional measure of intra-subject variablity. Safrit, Spray and Diewert (1980), in examining the theoretical distribution of VE, stated that VE may not be normally distributed, but failed to conclude that the distribution was definitely non-normal. One of the purposes of this study is to determine if the distribution of VE scores calculated from actual raw scores is non-normally distributed. If the empirical distribution is normal, many of the concerns of Schutz et al., (1973) and Safrit et al., (1980) will not be vital in analyzing VE data by an ANOVA.

Henry (1974) agreed with Schutz and Roy (1973) on the inadequacy of AE. While stating that CE and VE must always be looked at when interpreting performance error, he said that, at times, it may be necessary to use a composite score. Henry suggested using E^2 (where $E^2=CE^2+VE^2$) to which Schutz (1974) replied that E^2 is still a composite score and must be interpreted from CE and VE scores. Henry (1975), using multiple correlations, showed that E^2 was better than AE since the effect of VE is never excluded in E^2 while it may be in AE (when $CE/\sqrt{VE}>2.0$). Schutz (1979) conceded that, if a composite measure had to be used, then E^2 is preferable to AE but it still must be interpreted with respect to CE and VE.

Jones (1974) suggested that AE, not VE is the appropriate error score when the criterion is changed for each trial of a similar task. Roy (1974), replied that since KR is not given on every trial, the typical movement reproduction experiment is not a learning situation but a forgetting one. Roy argued that VE is a measure of forgetting and lack of consistency in performance which does not require the criterion for each trial to be similar in order to be interpreted.

Schmidt (1975) favored the use of AE claiming that for motor recall studies it is the preferable dependent measure for the following reasons: (a) the use of two dependent variables (CE and VE) may yield different results, thereby confusing any interpretation of results; (b) AE is the traditional measure, and (c) since the subject is required to minimize his error on each trial, AE is what should be measured. Schutz (1979) responded to each of these arguements, respectively, as such: (a) any theory should satisfy both performance dimensions as suggested by the CE and VE scores; (b) the fact that AE has been the traditional measure is sufficient reason to continue using it; and (c) since the purpose of the researcher is to <u>explain</u> performance, not only to <u>measure</u> it, CE and VE must be used in the interpretation.

In 1976, Newell stated that when one half of the subjects have positive CE's while the other half have negative CE's, the use of an average CE is inappropriate and AE should be used. In this situation Schutz (1979) agreed with Henry (1975) in that the absolute value of CE, $|CE|$, is the best measure.

The AE-CE-VE controversy then shifted from the theoretical

interpretations of these measures to more statistical ones. Roy (1976) stated that a good method of reporting all three error terms (AE or E, CE and VE) in studies is to analyze all three measures by a MANOVA since it controls for the Type I error rate. Roy provided a footnote which indicated that, based on work by Schutz and Roy, AE may be a linear composite of CE and VE and, therefore, a MANOVA could not be calculated. However, he stated that this rarely occurs across all subjects. Thomas (1977) replied that even though an absolute linear relationship between the three dependent variables may not exist, the problem of multicollinearity does. Multicollinearity has the effect of increasing the Type I error rate (Press, 1972). Thomas suggested analyzing VE and CE with a MANOVA and doing a separate ANOVA for AE or E. In replying to Thomas (1977), Roy (1977) agreed with the concept of multicollinearity but further complicated the issue by indicating that a high correlation may exist between CE and VE, thereby making a test of these variables by a MANOVA subject to the effects of multicollinearity. Safrit, Spray and Diewert (1980) caution against the use of all AE, CE and VE in a MANOVA for different reasons. An assumption in MANOVA designs is that the joint probability vector of the random vector be multivariately normally distributed. Safrit et al., showed that CE is marginally normal, but both VE and AE may be marginally non-normal, and concluded that until future empirical work shows that the violations of these assumptions are not serious, analyzing AE, CE and VE by a MANOVA should be avoided. Earlier work, however, has shown that the T^2 procedure is relatively robust to multivariate non-normality (Mardia, 1971).

Earlier, Thomas and Moon (1976) found AE scores to have higher reliabilities than VE and a greater number of significant differences were obtained with AE. These facts along with their finding that AE appeared to be more normally distributed about the target than VE allowed them to conclude that AE is the best dependent measure when conducting motor rhythm experiments.

Safrit et al., (1980) in stating that the distributions of AE and VE may be non-normal caution investigators in analyzing these dependent measures by an ANOVA. However, the violation of the normality assumption by itself is not serious (Boneau, 1960), but when interactive with violations of other assumptions, the Type I error rate is affected (Bradley, 1980). Therefore, if the researcher hesitates in using an ANOVA due solely to non-normality, he should check the other assumptions to see if they are satisfied.

While the area of which dependent measure is proper to use and report is obviously confusing, the following rule of thumb is generally accepted. Any investigator who can provide a logical explanation as to what information AE provides is justified in reporting it (Safrit et al., 1980).

Summary. As the wealth of literature has indicated, the choice of the dependent measure to be analyzed and interpreted is a subject of great controversy. Much of the debate deals with the conceptual interpretation of these measures, and thus is out of the range of the statistician, but a great deal of uncertainty surrounds the distributions and effects of using these dependent variables in an analysis of variance. Although many of the present problems will still exist, hopefully, the

question of the empirical distribution of VE and its subsequent effect on the Type I error rate will be adequately resolved at the conclusion of this study.

REFERENCES

Anderson, T. W. An Introduction to Multivariate Analysis. New York: Wiley, 1958.

Boneau, C. A. The effects of violations of assumptions underlying the t test. Psychological Bulletin, 1960, 57, 49-64.

Box, G. E. P. Problems in the analysis of growth and wear curves. Biometrics, 1950, 6, 362-389.

Box, G. E. P. Some theorems on quadratic forms applied in the study of analysis of variance problems. I. Effect of inequality of variance in the one-way classification. Annals of Mathematical Statistics, 1954a, 25, 290-302.

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 of correlation between errors in the two-way classification. Annals of Mathematical Statistics, 1954b, 25, 484-498.

Bradley, J. V. Nonrobustness in Z, t, and F tests at large sample sizes. Bulletin of the Psychonomic Society, 1980, 16, 333-336.

Burdick, J. A. Measurements of "Variability". Journal of General Psychology, 1972, 86, 201-206.

Collier, R. O., Baker, F. B., Mandeville, G. K., & Hayes, T. F. Estimates of test size for several test procedures based on conventional variance ratios in the repeated measures design. Psychometrica, 1967, 32, 339-353.

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

Gaito, J. Repeated measurements designs and counterbalancing. Psychological Bulletin, 1961, 58, 46-54.

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

Gaito, J. & Turner, E. D. Error terms in trend analysis. Psychological Bulletin, 1963, 60, 464-474.

Geisser, S. Multivariate analysis of variance for a special covariance case. Journal of the American Statistical Association, 1963, 58, 600-669.

Geisser, S. & Greenhouse, S. W. An extension of Box's results on the use of the F distribution in multivariate analysis. Annals of Mathematical Statistics, 1958, 29, 855-891.

Gnanadesikan, R. Methods for statistical data analysis of multivariate observations. New York: Wiley, 1977.

Goodman, D. & Schutz, R. W. DATASNIFF: A program to check data and compute summary descriptive statistics and correlations. Developed at the Physical Education Quantification Laboratory, University of British Columbia, 1975.

Greenwald, A. G. Within-subjects designs: To use or not to use? Psychological Bulletin, 1976, 83, 314-320.

Greenhouse, S. W. & Geisser, S. On methods in the analysis of profile data. Psychometrika, 1959, 24, 95-112.

Hakstian, A. R., Roed, J. C. & Lind, J. C. Two-sample T^2 procedure and the assumption of homogeneous covariance matrices. Psychological Bulletin, 1979, 86, 1255-1263.

Halm, J. UBC NORMAL. University of British Columbia Computing Center, 1970.

Henry, F. M. Variable and constant performance errors within a group of individuals. Journal of Motor Behavior, 1974, 6, 149-154.

Henry, F. M. Absolute error vs. "E" in target accuracy. Journal of Motor Behavior, 1975, 7, 227-228.

Holloway, L. N. & Dunn, O. J. The robustness of Hotelling's T^2. Journal of the American Statistical Association, 1967, 62, 124-136.

Hsu, P. L. Contributions to the theory of Student's t-test as

applied to the problem of two samples. Statistical Research Methods, 1938, 2, 1-24.

Huynh, H. Some approximate tests for repeated measurements designs. Psychometrika, 1978, 48, 161-175.

Huynh, H. & Feldt, L. Conditions under which mean square ratios in repeated measurements designs have exact F-distributions. Journal of the American Statistical Association, 1970, 65, 1582-1585.

Huynh, H. & Feldt, L. S. Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. Journal of Educational Statistics, 1976, 1, 69-82.

Ito, K. On the effect of heteroscedasticity and non-normality upon some multivariate test procedures. In P. R. Krishnaiah (Ed.), Multivariate analysis-II. New York: Academic Press, 1969.

Ito, K. & Schull, W. J. On the robustness of the T^2 test in multivariate analysis of variance when variance-covariance matrices are not equal. Biometrika, 1964, 51, 71-82.

Jones, B. "What is the best measure of accuracy of movement duplication?" Unpublished paper, University of Waterloo, 1974.

Kogan, L. S. Analysis of variance: Repeated measurements. Psychological Bulletin, 1948, 45, 131-143.

Laabs, G. J. Retention characteristics of different reproduction cues in motor short-term memory. Journal of Experimental Psychology, 1973, 100, 168-177.

Lana, R. E., & Lubin, A. The effect of correlation on the repeated measures design. Educational and Psychological Measurement, 1963, 23, 729-739.

Mardia, K. V. The effect of nonnormality on some multivariate tests and robustness to nonnormality in the linear model. Biometrika, 1971, 58, 105-121.

Mauchly, J. W. Significance test for sphericity of a normal n-variate distribution. The Annals of Mathematical Statistics,

1940, 29, 204-209.

Mendoza, J. L., Toothaker, L.E. & Nicewander, W.A. A Monte Carlo comparison of the univariate and multivariate methods for the groups by repeated measures design. Multivariate Behavioral Research, 1974, 9, 165-178.

Morrow, J. R. & Frankiewicz, R. G. Strategies for the analysis of repeated and multiple measures designs. Research Quarterly, 1979, 50, 297-304.

Newell, K. M. More on absolute error, etc. Journal of Motor Behavior, 1976, 8, 139-142.

Press, S. J. Applied multivariate analysis. New York: Holt, Rinehart, and Winston, 1972.

Rogan, J. C. , Keselman, H. J. & Mendoza, J. L. Analysis of repeated measurements. British Journal of Mathematical and Statistical Psychology, 1979, 32, 269-286.

Rouanet, H. & Lepine, D. Comparison between treatments in a repeated-measures design: ANOVA and multivariate methods. The British Journal of Mathematical and Statistical Psychology, 1970, 23, 147-163.

Roy, E. A. "Measures of accuracy of movement duplication - A reply to Jones." Unpublished paper, University of Waterloo, 1974.

Roy, E. A. Measuring change in motor memory. Journal of Motor Behavior, 1976, 8, 283-287.

Roy, E. A. Measuring change: A reply to Thomas. Journal of Motor Behavior, 1977, 9, 255-256.

Schmidt, R. A. Critque of Henry's paper. In L. E. Smith (Ed.), Psychology of motor learning. Chicago: Athletic Institute, 1970.

Schmidt, R. A. A schema theory of discrete motor learning. Psychological Review, 1975, 82, 225-260.

Safrit, M. J., Spray, A. J. & Diewert, G. L. Methodological

issues in short-term motor memory research. Journal of Motor Behavior, 1980, 12, 13-28.

Scheffe, H. The Analysis of Variance. New York: Wiley, 1959.

Schutz, R. W. Absolute error. Journal of Motor Behavior, 1974, 6, 299-301.

Schutz, R. W. Specific problems in the measurement of change: Longitudinal studies, difference scores, and multivariate analyses. In D. Landers & R. Christina (Eds.), Psychology of Motor Behavior and Sport - 1977, Champaign, Illinois: Human Kinetics, 1978.

Schutz, R. W. Absolute, Constant, and Variable Error: Problems and solutions. In D. Mood (Ed.), Proceedings of the Colorado Measurement Symposium. University of Colorado Press: Boulder, Colorado, 1979, 82-100.

Schutz, R. W. & Gessaroli, M. E. The effects of block size and heterogeneity of covariance on Type I error rates with Constant error and Variable error data. in Psychology of Motor Behavior and Sport - 1979. Champaign, Illinois: Human Kinetics, 1980, 633-642.

Schutz, R. W., Goodman, D. & Roy, E. A. "More on error." Paper presented at the First Canadian Congress for the Multidisciplinary Study of Sport and Physical Activity, Montreal, Quebec, October, 1973.

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

Stoloff, P. H. Correcting for heterogeneity of covariance for repeated measures designs of the analysis of variance. Educational and Psychological Measurement, 1970, 30, 909-924.

Thomas, J. R. & Moon, D. H. Measuring motor rhythmic ability. Research Quarterly, 1976, 47, 20-32.

Thomas, J. R. A note concerning analysis of error scores from motor memory research. Journal of Motor Behavior, 1977, 9, 251-253.

Wilson, A. & Lange, D. E. Univariate analysis of variance for

repeated measures. The Journal of Experimental Education, 1972, 40, 83-85.

Wilson, K. The sampling distributions of conventional, conservative and corrected F-ratios in repeated measurement designs with heterogeneity of covariance. Journal of Statistical Computing and Simulation, 1975, 3, 201-215.

Woodworth, R. S. Experimental psychology. New York: Holt, 1938.