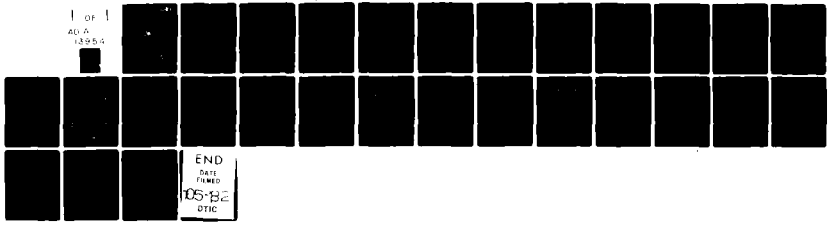
AD-A113 954 NAVAL BIODYNAMICS LAB NEW ORLEANS LA F/G 6/19
REPEATED MEASURES OF HUMAN PERFORMANCE: A BAG OF RESEARCH TOOLS--ETC(U)
NOV 81 A C BITTNER, R C CARTER
UNCLASSIFIED NBDL-81R011 NL



NBDL - 81R011

30

REPEATED MEASURES OF HUMAN PERFORMANCE: A BAG OF RESEARCH TOOLS

ALVAH C. BITTNER, JR. and ROBERT C. CARTER, JR.

AD A113954

13 NOVEMBER 1981

NAVAL BIODYNAMICS LABORATORY
New Orleans, Louisiana

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NBDL - 81R011	2. GOVT ACCESSION NO. AD-A113 954	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* Repeated Measures of Human Performance: A Bag of Research Tools		5. TYPE OF REPORT & PERIOD COVERED Research Report
		6. PERFORMING ORG. REPORT NUMBER NBDL - 81R011
7. AUTHOR(s) Alvah C. Bittner, Jr., and Robert C. Carter		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Biodynamics Laboratory Box 29407 New Orleans, LA 70189		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS MF.58.524-002-5027
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Medical Research and Development Command Bethesda, MD 20014		12. REPORT DATE November 1981
		13. NUMBER OF PAGES 24
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*		15. SECURITY CLASS. *(of this report)* Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Repeated measures, statistical criteria, human performance tests, methodology, intervention experiments, environmental research paradigms

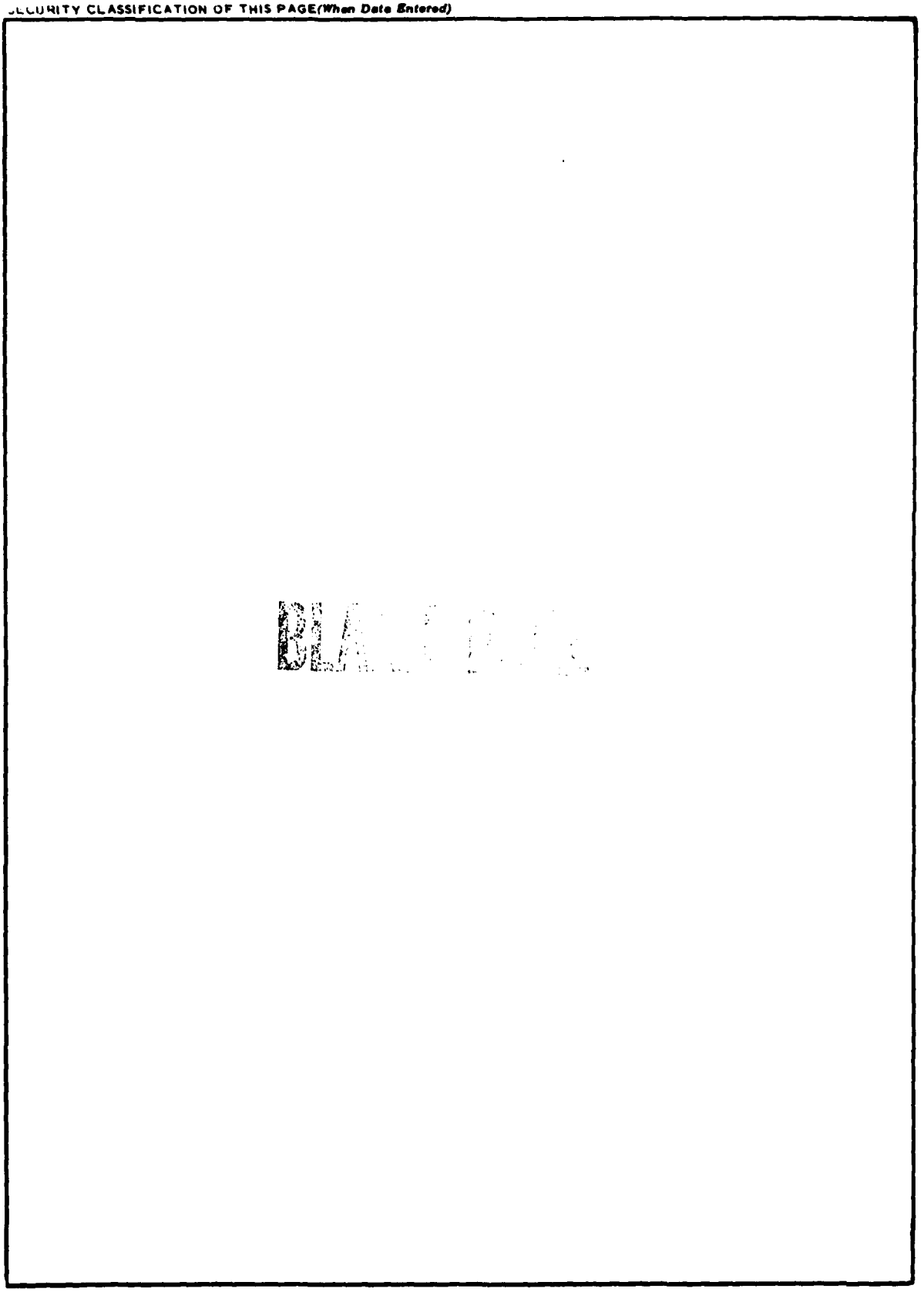
20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

Research tools are described which are applicable to repeated measures of human performance. In the first section, statistical criteria for tasks are delineated, tools for assessment are described, and examples of applications are given. In the second section, multiple subject and single subject analyses of intervention experiments are considered with major focus on the methodological tools. The final section summarizes these tools with examples of their application.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601 |

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*


BLANK

NBDL - 81R011

REPEATED MEASURES OF HUMAN PERFORMANCE: A BAG OF RESEARCH TOOLS

Alvah C. Bittner, Jr. and Robert C. Carter

Bureau of Medicine and Surgery
Work Unit No. MF58.524-002-5027

Approved by

Channing L. Ewing, M. D.
Chief Scientist

Released by

Captain J. E. Wenger MC USN
Commanding Officer

Naval Biodynamics Laboratory
Box 29407
New Orleans, LA 70189

SUMMARY

THE PROBLEM

Human performance test methodologies for use in environmental research are being developed at the Naval Biodynamics Laboratory. Repeated measures on the same subjects are used almost exclusively in this and many other intervention studies (e.g., drug and clinical). Suitable tasks, experimental paradigms and statistical tools are required to insure the value of repeated measure investigations.

FINDINGS

Research tools are described which are applicable to repeated measures of human performance. In the first section, statistical criteria for tasks are delineated, tools for assessment are described, and examples of applications are given. In the second section, multiple subject and single subject analyses of intervention experiments are considered with major focus on the methodological tools. The final section summarizes these tools with examples of their application.

ACKNOWLEDGEMENTS

This paper will be published in J. C. Guignard (Ed.) Proceedings of the International Workshop on Research in Human Motion and Vibration Studies, New Orleans, 16-18 September 1981. Requests for reprints may be sent to Dr. Alvah C. Bittner, Jr., at the Naval Biodynamics Laboratory.

The authors would like to acknowledge the contribution of CDR Robert S. Kennedy, USN, MC (Ret.) to this research effort.

Trade names of materials or products of commercial or non-government organizations are cited where essential for precision. Their use does not constitute official endorsement or the approval of the use of such hardware or software.

INTRODUCTION

Investigations of vibration and other environmental effects almost exclusively employ repeated measures of subjects according to Kennedy and Bittner (1977). The general approach in such studies is to collect data on one or more trials conducted Before (B), During (D) and After (A) exposure. Evaluation of the suitability of tasks for repeated measurements and analysis methodologies will be the concern of this report.

Selection of a repeated measures paradigm follows from both theoretical and practical considerations. First, interest in the time course of development of and recovery from environmental effects frequently dictates repeated measures. Time-course measurements of a single individual or team during an environmental experiment, for example, may be expected to reveal features of response to environmental change which would not be observed if a composite of several individuals, each measured at different times were employed (Estes, 1956). In addition, it would be impracticable to study the time-course of effects with independent groups due to the prohibitive numbers of subjects which would be required. Other reasons for advocating the use of repeated measures are the increased measurement sensitivity and economical features of such experiments (Fisher, 1935, 1966; Sutcliffe, 1980; Winer, 1971). Individual differences in subjects may be removed under appropriate repeated measures designs, but remain part of the "error" in independent groups designs. Figure 1 is a nomogram which illustrates the impact of sample size (N) and correlaton between measures on the minimum significant ($p \geq .05$) differences (D) for one and two-tailed tests (Carter, Kennedy, & Bittner, 1981). Measures for independent groups, by definition, would have an expected $R = 0$; while, repeated measures would generally yield $R > 0$. Assuming a fixed N, the change in "sensitivity (D)" with increasing R from independence ($R = 0$) to complete dependence ($R = 1$) can be seen to be quite large. Similarly, with D fixed, the "economy" of repeated measures can be seen by noting the reductions in N for repeated ($R > 0$) verses independent ($R = 0$) groups. The last and often most potent argument for repeated measures is the requirement to reduce subject risk in hazardous environments. Increased economy through use of repeated measures implies reduced subject risk with fewer subjects and numbers of exposures required for a given level of sensitivity in addition to reduced financial costs. The reduction of subject risk and other considerations have led to the adoption of repeated measures experimentation in this laboratory.

The requirement for Before-During-After (BDA) experimentation has motivated this laboratory to develop applicable tasks and methodologies. One project is underway to evaluate performance test suitability for repeated measures applications (Kennedy & Bittner, 1977; Carter, Kennedy, & Bittner,[1] 1980; Kennedy, Carter, & Bittner, 1980; Shannon, Carter, & Boudreau, 1981). A second project, focusing on the application of Box-Jenkins (1970) Time-Series methodology to BDA experiments, is nearing completion; findings from this program have already evidenced considerable promise for this approach (Carter, 1980; Glass, Wilson, & Gottman, 1975). In addition to these projects, others are underway which are directed primarily at the effects of impact, vibration,

---

[1] This was identified as the Performance Evaluation Tests for Environmental Research (PETER) Program in earlier reports.

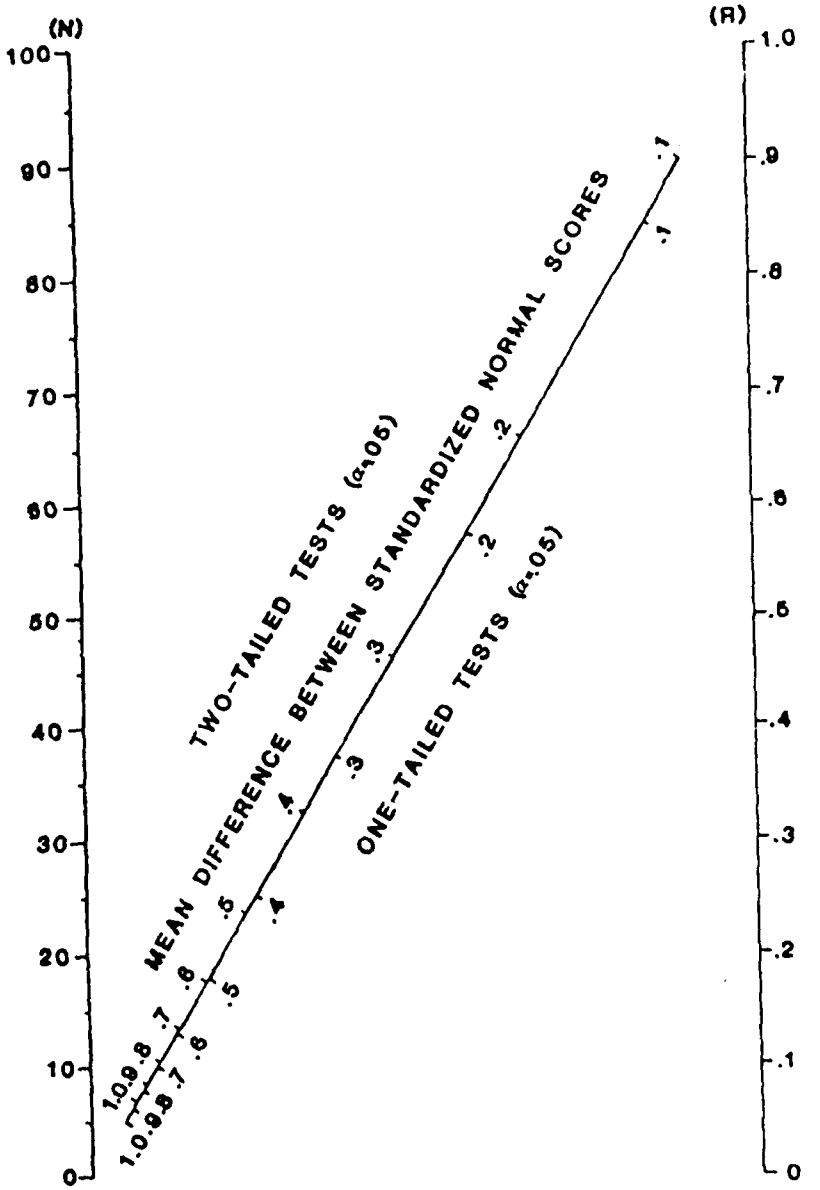


Figure 1. Nomogram Relating Sample Size (N), Intertrial Correlation (R), and the Smallest Significant (p = .05) Difference (D)

(Carter, Kennedy, & Bittner, 1981)

and motion sickness on performance. Requirements for environmental research, as part of these programs, has driven developments involving BDA design and analysis. For example, recent vibration experiments have been conducted with an aim to developing a methodology applicable to long-term investigations (Guignard, Bittner, & Carter, 1981). Altogether, this laboratory's programatic efforts have resulted in the assembly of a "bag of research tools" which are of value for repeated measures investigations.

The purposes of this report are twofold. In the first section, statistical criteria for tasks to be repeatedly measured are delineated, and examples are given of desirable and undesirable tasks. The second section focuses on analysis of intervention experiments including both multiple subject experiments and single subject analysis. The last section summarizes the "bag of research tools" described in the earlier sections.

TASK SELECTION FOR REPEATED MEASURES

Statistical Criteria

Candidate tests for repeated measures studies should meet rigorous statistical qualification (Jones, 1972, 1980; Kennedy & Bittner, 1977; Kennedy, et al., 1980). Meaningful repeated measures, as outlined by Jones (1972, 1980), generally require that means, variances, and intertrial correlations are "well-behaved" when obtained under constant (baseline) conditions. Baseline conditions, identified by Kennedy and Bittner (1977, 1980) for performance tests, typically involve daily administration of tasks to (15-20) subjects for 15 workdays.[2] Assessment of tasks across days permits assessment of task differential changes with practice, which are uncontaminated by within-day autocorrelative effects (Campbell & Stanley, 1966; Thorndike, 1949). Unambiguous assessment of differential change with practice was deemed necessary because of the substantial evidence for such change (cf., Alvares & Hulin, 1972). When such changes are occurring, it is difficult to establish "what is being measured" and to make scientific generalizations (Bittner, 1979; Jones, Kennedy, & Bittner, 1981). Specific baseline condition statistical characteristics which are considered necessary are described below.

Means. The criterion for means is that they change in a linear manner or are unchanging over trials. This criterion has been identified by Campbell and Stanley (1966) as a requirement for interpretation of repeated measures results. Significantly, it is unnecessary that this criterion be met from the first trial, if practice is carried out beyond a point where it is obtained before beginning a cycle of BDA. Such a point in practice, it is noteworthy, is expected with sufficient practice (Reynolds, 1952; Fitts & Posner, 1967). Hence in task evaluations, means are tested sequentially, dropping leading days, until this criterion of linearity is met.

Statistical techniques for accomplishing means analysis include graphical, analysis of variance (ANOVA), and orthogonal polynomial analyses. The BMDP2V (Dixon & Brown, 1977) computer program, with option ORTHOGONAL, provides a direct and rigorous analysis.

---

[2] The need for other supplementary baseline (e.g., within day) investigations was noted but not developed by these authors.

Variances. The criterion for within-trial variances is that they are homogeneous over trials. This criterion, in addition to constant intertrial correlations, constitutes compound-symmetry, the traditional assumption for simple repeated measures ANOVA (Box, 1950; Scheffé, 1959; Winer, 1971). As with the means, it is unnecessary that this criterion be met from the first trial if practice is carried out beyond the point where the criterion is obtained. Thus, in task evaluation, variances also are tested sequentially, dropping leading days, until the criterion is met. Statistical techniques for accomplishing this analysis include graphical and a multitude of analytic tests. Where the normality assumption holds, familiar statistical analyses (e.g., Fmax) may be employed for this purpose; these are extremely sensitive to nonnormality (Scheffé, 1959, Chapter 10). Alternate analyses are suggested where normality is questionable, including, Scheffé's (1959) log-transformed variance or related Miller's Jackknife analyses (Hollander & Wolfe, 1973). The exact procedure for establishing homogeneity of variance is less important than its unambiguous establishment.

Correlations. The criterion for the cross-day correlations is that they are differentially stable (constant). As with the criteria for homogeneous variances described above, the differential stability criterion is embedded in the traditional (or compound symmetry) requirement for simple repeated measures ANOVA. Differential stability and homogeneity of variance, in addition to their implications for ANOVA, are sufficient indications that the Spearman-Brown Formula may be applied to estimating the reliability of a test with changes in test length (Thorndike, 1949; Winer, 1971). Figure 2 (Kennedy et al. 1980) shows the tradeoff of reliability and time; it provides a method of assessing the length of testing required for a reliability found desirable from consideration of Figure 1. Differential stability, most importantly, implies that the same attribute is being measured on each occasion of measurement. With attribute changes, statistical testing may be possible, but attribution of effect and scientific generalization are precluded (Jones, et al., 1981).

Statistical tests for differential stability have been of continuing concern. In an earlier paper, Bittner (1979) reviewed and illustrated graphical and analytical methods which were applied in early task investigations. More recently, the method of Steiger (1980a, 1980b) has been routinely applied for stability determination. Other methods which have captured interest include possible applications of factor analysis, nonparametric directional tests, and jackknife approaches (e.g., Jöreskog, 1969; Shannon, 1980; Jones, 1981; Gnanadesikan, 1977). However, because of the omnibus character of the Steiger analysis and its computer implementation, it has continued to be recommended. It has been possible to test sequentially for differential stability by manually dropping leading days. This procedure, it is noteworthy, was supported by early work of Jones (1970a, 1970b, 1972) in which differential stability was found to emerge with practice. The recent development of a nonmanual stepwise program for Steiger (1980a, 1980b) analysis gives added support for the standard use of this analysis.[3]

Overall, the task criteria described above lead to straightforward experimental design, simplicity of statistical analysis and unambigious interpretation of results. Augmentation of these criteria with others

---

[3] Regarding this computer program, LCDR Robert C. Carter can be contacted at the Naval Biodynamics Laboratory, Box 29407, New Orleans, LA. 70189.

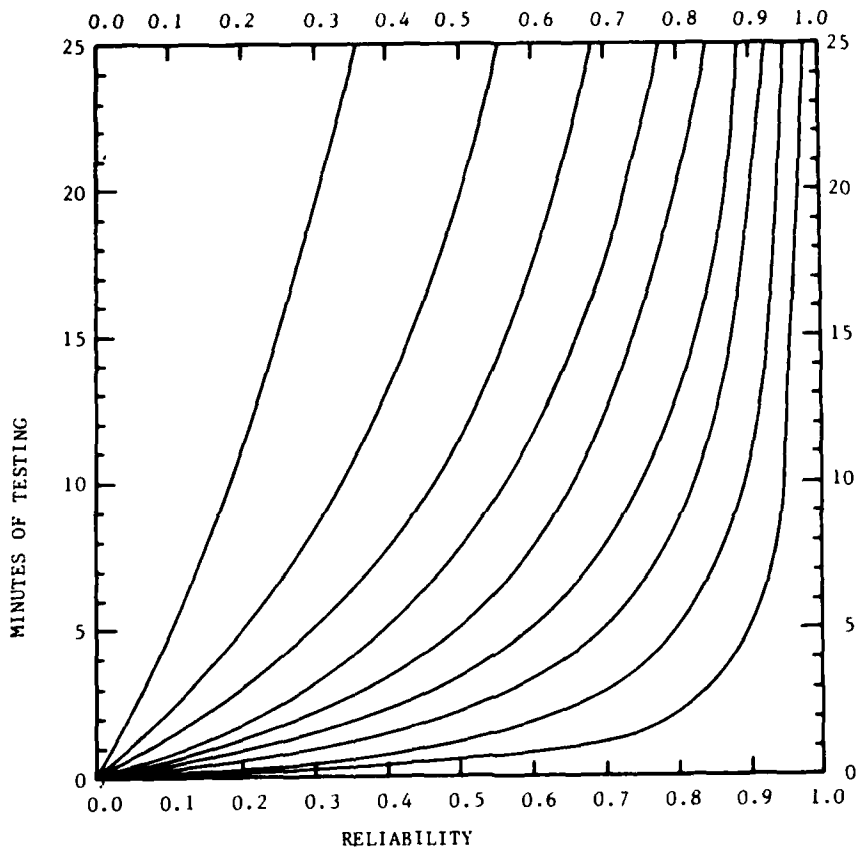


Figure 2. Tradeoffs between Intertrial Correlation and Test Time (Kennedy, Carter, & Bittner, 1980).

may be anticipated for applications within days, where autocorrelative effects may be anticipated (c.f., Thorndike, 1949; Campbell & Stanley, 1966). In particular, investigations employing Box-Jenkins (1970) models may be required; they will be described as part of the second section of this report. Pertinently, the baseline criteria described above also support the assumptions required for within day investigations. Tasks which have been evaluated using the statistical criteria are summarized in other reports (Kennedy, et al., 1980; Kennedy, 1981).

Tasks Evaluation Examples

Two evaluations of the statistical suitability of tasks will be given to illustrate applications of the above criteria. The first evaluation is of the Spoke Control Task, a motor dexterity task which was considered as part of a larger investigation (Bittner, Lundy, Kennedy, & Harbeson, in press). This task, which successfully met the statistical criteria given above, was recently employed in an investigation of vibration effects (Guignard, et al., 1981). The second example is a time estimation measure which has been shown to be unsuitable for repeated measures applications (McCauley, Kennedy, & Bittner, 1980). Together these examples give illustrations of task success and failure.

Spoke Task. Computer generated paper-and-pencil Control Task (CT) forms were produced and printed by a programmed WANG™ Computer on unlined display sheets. The display sheets (43 cm x 28 cm) contained 32 circular targets arranged concentrically around a central circular target (marked 0). Each target was 9.5 mm in diameter and located 120.6 mm from the central target. Distance from the center of one target to an adjacent target was 25.4 mm. The number "1" was in the twelve o'clock position and began an ascending sequence in a clock-wise direction. Each of 18 enlisted male volunteers was required alternately to tap his stylus on the center target (0) and on each of the numbered circles (1,2,...,32) in succession (0,1; 0,2; ...; 0,32). Errors, if any, were corrected as they were observed. The CT score was the time to completion as measured by a stop-watch. Subjects were tested daily for 15 consecutive workdays Monday through Friday between 0800 and 1000.

Figure 3 shows the means and standard deviations for the CT over days. A slow linear decline in the means is suggested, but no change is seen in the standard deviations. The overall change in means was confirmed by analysis of variance (ANOVA) with $F(14, 238) = 3.54$; $p < .01$. Of the overall sums of squares, 55% was accounted for by a very highly significant linear component, $F(1,238) = 27.3$ ($p < 10^{-5}$), with no significant indication of higher order components, $F(13, 238) = 1.7$ ($p > .06$). The apparent lack of change in standard deviations was also confirmed by a nonsignificant $F\text{max}(15,17) = 2.92$ ($p > .1$). Hence, the CT means and variances were stable from the first day.

Table 1 contains the CT reliability coefficients across all days from which the correlation traces in Figure 4 were drawn. The traces shown in this figure were drawn for selected Base Days (1, 2, 4, 8, 10, and 12) by left justifying the appropriate row of the correlation matrix in terms of days after base performance (Bittner, 1979). Examining Figure 4, it can be noted that subsequent to Base Day 1, the traces are level and

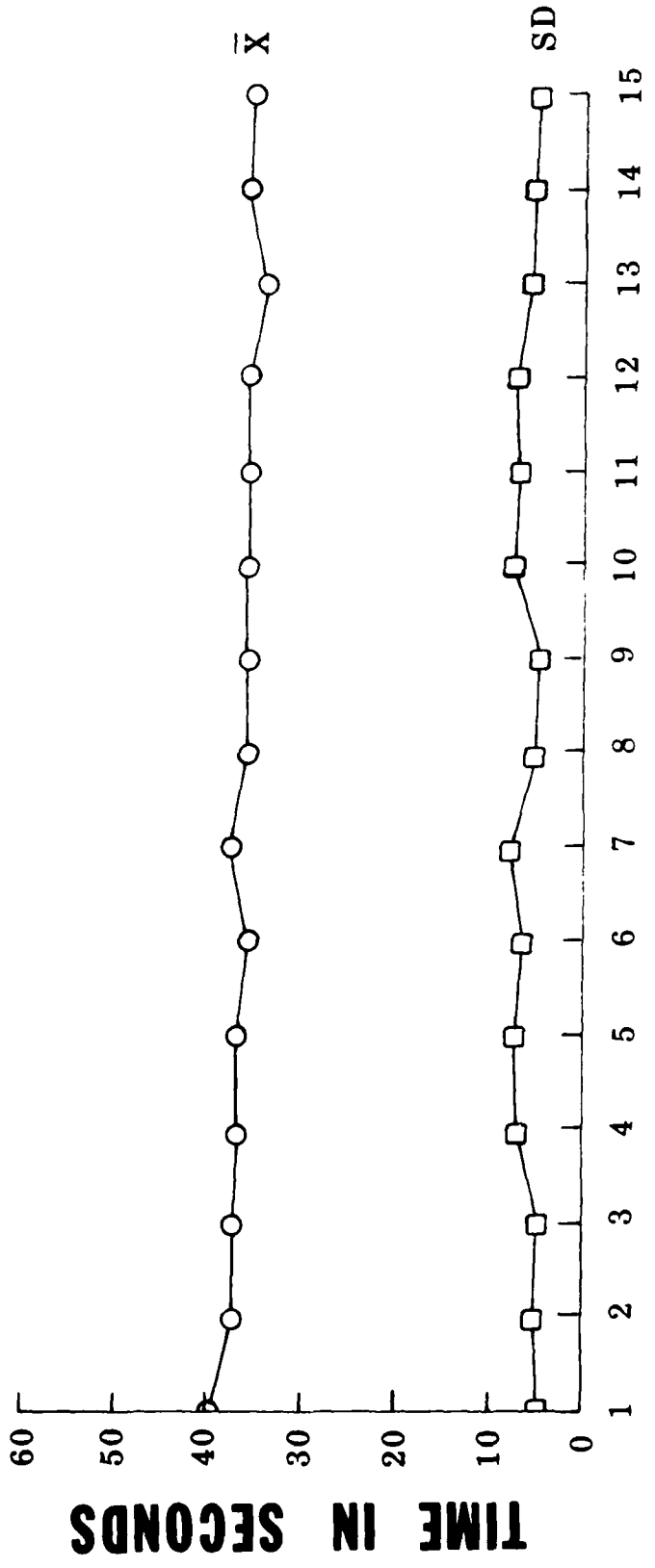


Figure 3. Spoke Control Task (CT) Means and Standard Deviations over 15 Days (N = 18)

(Bittner, Lundy, Kennedy, & Harbeson, in press)

Table 1: SPOKE CONTROL TASK RELIABILITIES OVER 15 DAYS (N = 18)
(Bittner, Lundy, Kennedy, & Harbeson, in press)

DAYS	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		.72	.77	.69	.71	.61	.65	.65	.69	.69	.71	.73	.57	.63	.75
2			.88	.87	.85	.85	.72	.86	.79	.84	.85	.85	.86	.79	.92
3				.87	.82	.79	.72	.79	.67	.84	.70	.77	.67	.69	.88
4					.85	.81	.70	.82	.76	.85	.77	.86	.81	.81	.90
5						.85	.83	.85	.81	.91	.89	.91	.84	.61	.88
6							.89	.87	.89	.84	.91	.85	.78	.77	.89
7								.78	.84	.78	.83	.80	.69	.58	.80
8									.88	.86	.86	.81	.83	.79	.90
9										.83	.88	.86	.79	.76	.84
10											.82	.88	.85	.64	.91
11												.90	.82	.77	.87
12													.86	.75	.88
13														.68	.83
14															.77

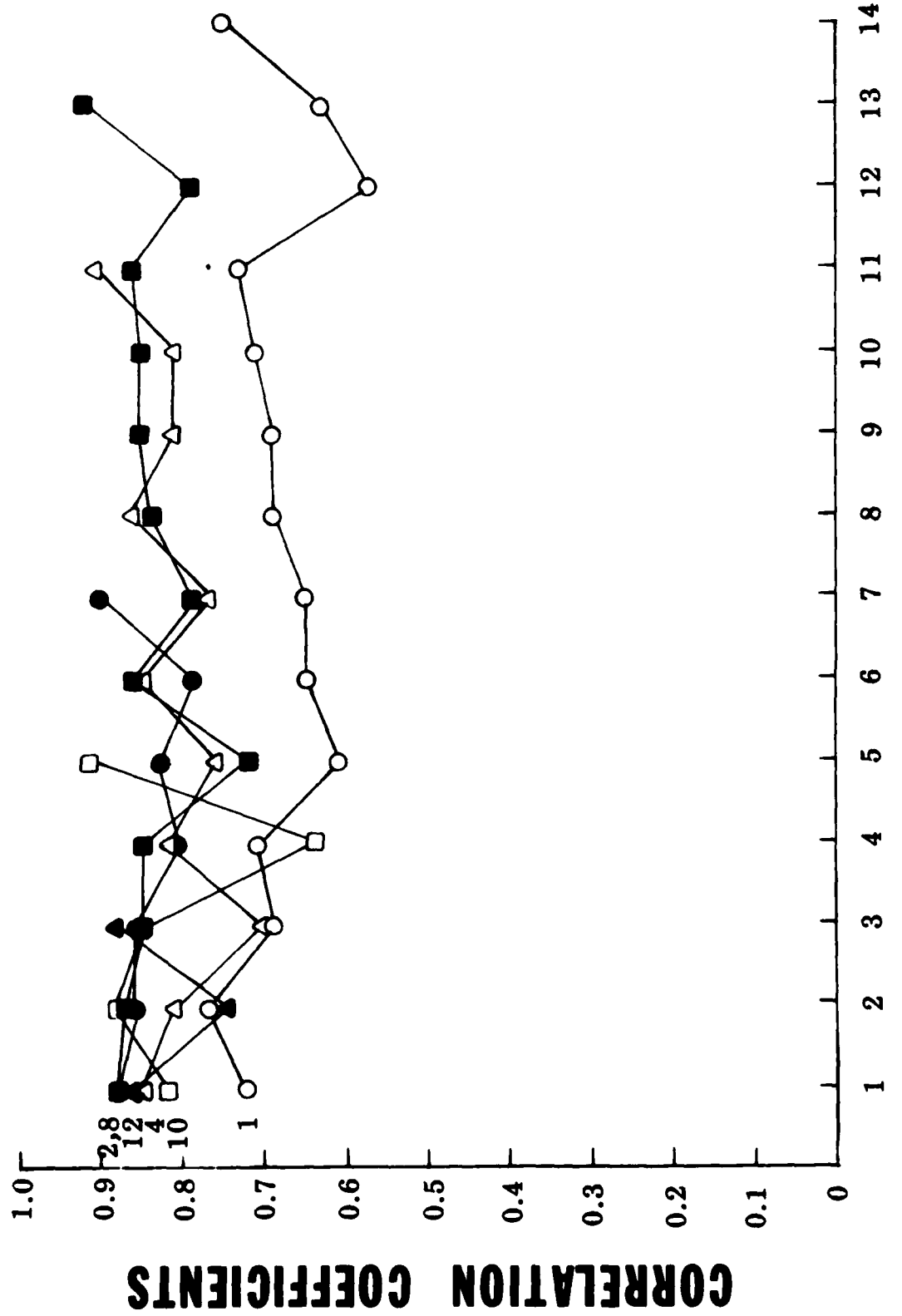


Figure 4. Comparison Spoke Control Task (CT) Reliabilities between Selected Base Days (1, 2, 4, 8, 10, 12) and Those Following over 15 Days (Bittner, Lundy, Kennedy, & Harbeson, in press).

overlapping. This pattern indicates that reliabilities are differentially stable subsequent to the first session (Bittner, 1979; Jones, 1980). A statistical test of differential-stability using the approach of Steiger (1980a, 1980b), however, yielded $\chi^2(104) = 103.8$ ($p > .49$) which indicated a constant ($\bar{r} = .799$) correlation even from the first session. Conservatively, the CT is both differentially stable and has high task definition subsequent to the first session.

Time Estimation. Constant Error(CE), a global estimation measure, was considered as part of the larger McCauley, et al. (1980) investigation. Daily scores for each of 19 enlisted men were derived from 5 productions of 8 time intervals (2, 3, 5, 6, 8, 9, 11 and 12 seconds) without the subjects' knowledge of results. On each of 15 weekdays, the 40 trials (8 intervals by 5 replications) were given in random order. A subject's daily CE was his mean deviation from the specified intervals over trials.

CE means, variances and cross day correlations were analyzed subsequent to data collection. Figure 5 shows the means and standard deviations and suggests little change with practice. A repeated measures ANOVA, it is noteworthy, yielded $F(14,252) = 0.90$ ($p > .56$) for Days. The cross day correlation results, given in Table 2 and illustrated in Figure 6, are dramatically different. Examining this figure, it can be noted that the reliability across Day 1 and Day 2 is 0.80 but that the reliabilities between Day 1 and succeeding days falls effectively to zero. The average reliability between immediately adjacent days ($\bar{r}_{1,2}$, $\bar{r}_{2,3}$, ..., $\bar{r}_{14,15}$) can be computed from Table 1 to also be $\bar{r} = 0.80$. However, as seen in Figure 3 the fall-off pattern with succeeding days continues and can be seen as late as Day 12. Even if stable beyond this point, the more than three hours practice required would make this task unattractive for repeated measures research. McCauley, et al. (1980) also found such instability for a variety of other time estimation global measures, transformed measures, and subtask scores. Certainly the results did not contradict Posner's (1978) view that there is no general time estimation trait.

ANALYSIS OF REPEATED-MEASURES EXPERIMENTS

This section of the paper describes some tools for design and analysis of repeated-measures experiments and techniques for multiple-subject and single-subject experiments.

Multiple-Subject Experiments

The most commonly analyzed effect of motion and vibration is a change of mean performance. In an experiment which includes measurements Before (B), During (D), and After (A) the treatment with no carry over of treatment into A, the contrast $(\bar{D} - (\bar{B} + \bar{A})/2)$ represents the mean effect of the treatment, independent of the mean effect of practice which is represented by $\bar{A} - \bar{B}$. Figure 7 illuminates these constructs which respectively are identical with the quadratic and linear orthogonal polynomials for trends in the three repeated measures B, D, and A. The BMD2V (Dixon & Brown, 1977) computer program, described earlier, may be used to calculate statistical tests of these contrasts and their interactions with other contrasts yet to be discussed.

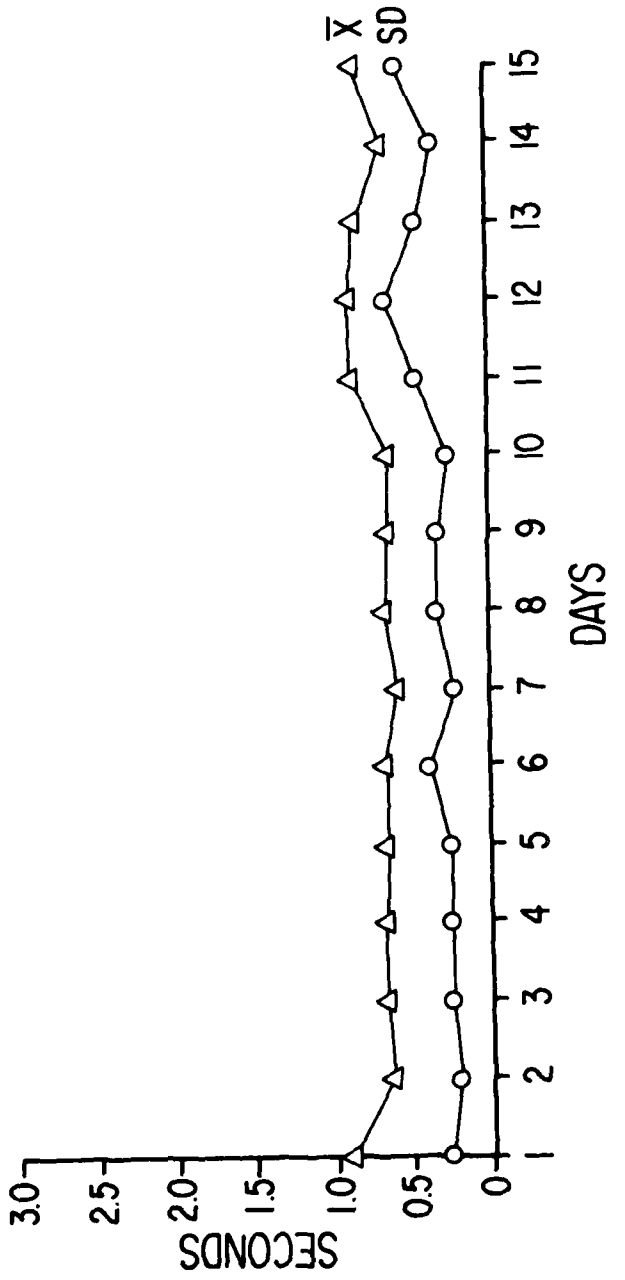


Figure 5. Constant Error (CE) Means and Standard Deviations over 15 Days (N = 19)

(McCauley, Kennedy, & Bittner, 1980).

Table 2: CONSTANT ERROR (CE) RELIABILITIES OVER 15 DAYS (N = 19)
(McCauley, Kennedy, & Bittner, 1980)

DAYS	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		.80	.40	-.14	.08	-.04	.16	.08	.03	-.12	-.19	-.05	-.21	-.26	-.26
2			.59	.22	.34	.28	.44	.40	.30	.14	.07	.16	-.05	-.02	-.07
3				.67	.73	.49	.54	.37	.20	.09	.12	.16	.12	.06	.03
4					.70	.69	.65	.53	.38	.28	.25	.27	.28	.19	.12
5						.80	.65	.62	.55	.38	.32	.42	.37	.36	.28
6							.83	.87	.82	.63	.57	.57	.52	.55	.37
7								.79	.70	.61	.53	.61	.53	.46	.39
8									.94	.80	.75	.72	.57	.66	.47
9										.84	.73	.72	.54	.62	.46
10											.76	.90	.82	.78	.78
11												.75	.61	.70	.54
12													.88	.84	.83
13														.89	.96
14															.90

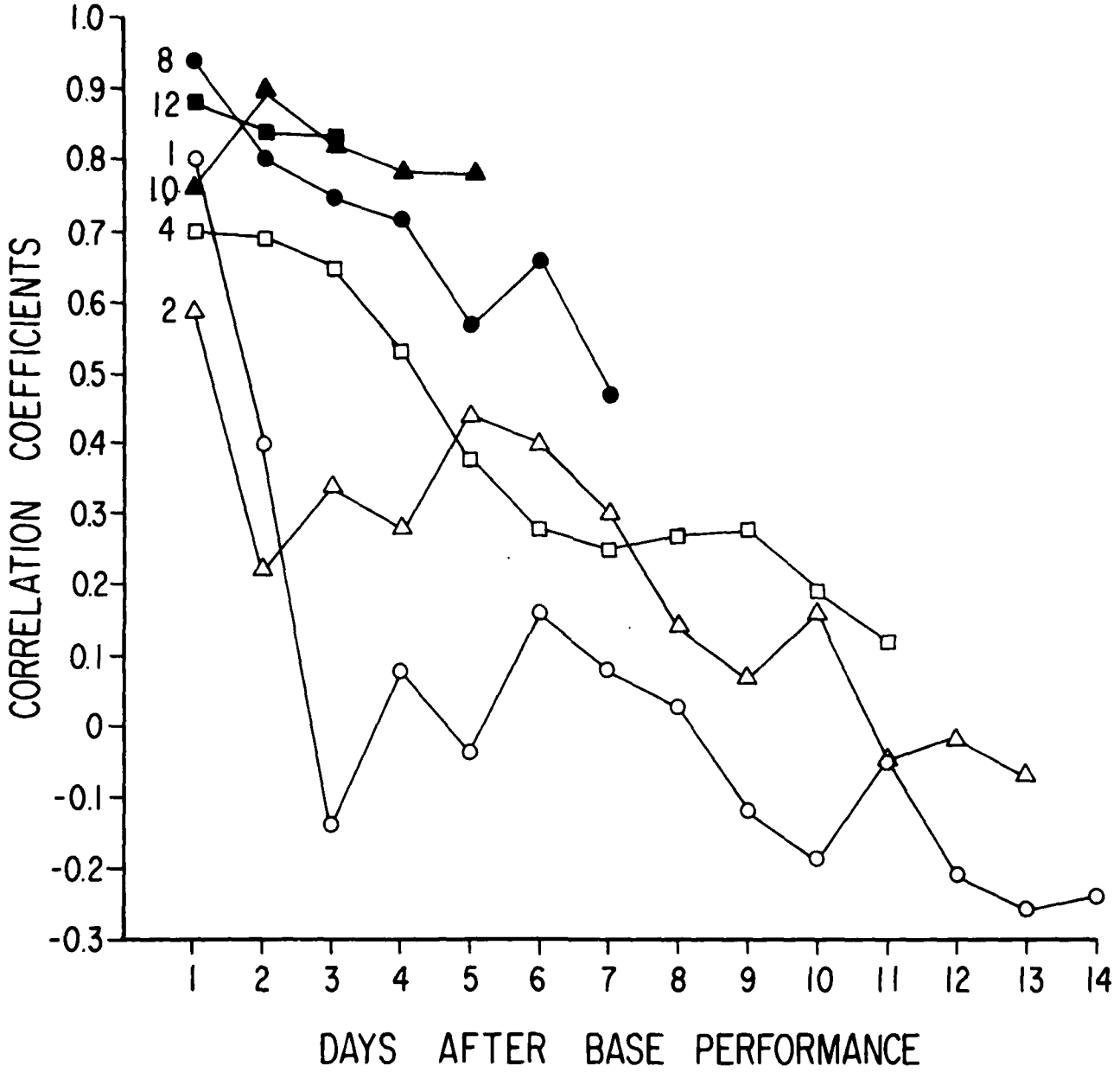


Figure 6. Comparison of Constant Error (CE) Reliabilities between Selected Base Days (1, 2, 4, 8, 10, 12) and Those Following (McCauley, Kennedy, & Bittner, 1980).

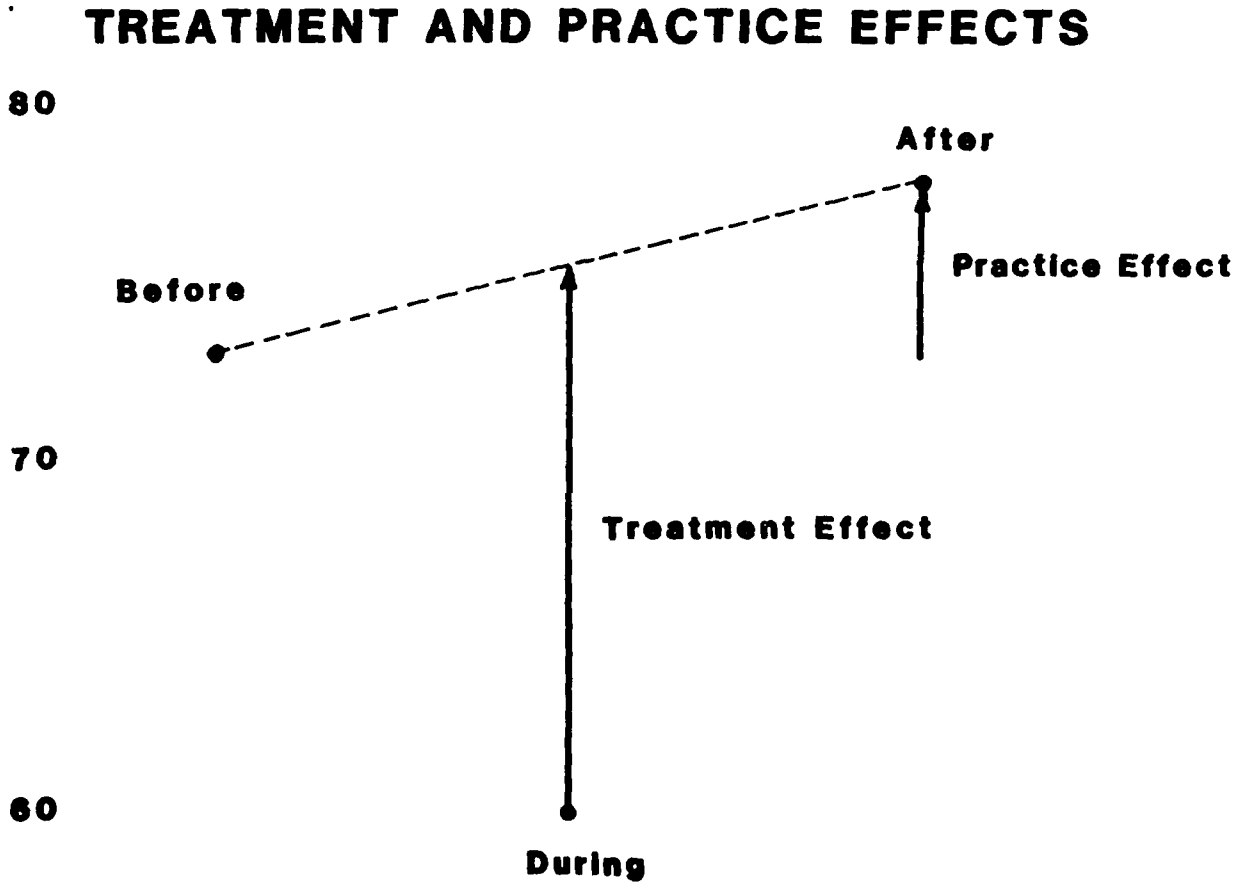


Figure 7. Example of Treatment and Practice Effects.

If a second experiment is conducted employing the same subjects to produce measurements B', D', and A' then the treatment and practice mean effects in the two experiments can be compared directly. This comparison may be made because the treatment and practice contrasts are independent of the level of performance B + D + A or B' + D' + A', which would tend to increase with practice from the first to the second experiment. In Figure 8 the independence of the contrasts (D - (B + A)/2) can be seen across four sequential experiments at 8, 16, 32, and 8 Hz vibration conditions. The 8 Hz conditions show statistically consistant decrements in the first and last experiments while other conditions show no effects. These data were obtained in a successful application of this approach where experiments were typically separated by intervals of several weeks (Guignard, et al., 1981).

In addition to effects on the means, the treatments may also affect the variances as illustrated in Figure 9. Changes of variances should be considered both for behavioral interpretations and for validation of assumptions underlying the analyses of effects on means. In terms of behavior, variances during the treatment may decrease if the treatment causes the subjects to adopt a stereotyped response, or prevents them from responding. Variances may increase if the treatment affects subjects to varying degrees. Other phenomena may also alter the variance of performance, and any inhomogeneity of variance raises questions about the validity of many techniques for assessing mean effects. The literature of statistics abounds with tools for comparing variances; at least one of them should be in an experimenter's tool bag.

Intertrial correlations should be examined for evidence of changes in the performance standings of the subjects relative to each other. Changes of the correlations can be tested with Steiger's MULTICORR computer program (1980a, 1980b). If correlations between treatment (D) and baseline (A and B) scores are lower than correlations between baselines, then subjects were not all equally affected by the treatment, nor was the effect linearly related to baseline scores. Figure 10 gives a hypothetical example of changes in correlations with environmental impact. These results would be expected if the treatment disrupts the abilities typically employed on a task so that subjects alter their test-taking strategy. In general, intertrial correlations represent the degree of consistency in subjects' responses to the treatment. If the correlations change, then the experimenter is alerted to an inconsistant effect.

Even if the intertrial correlations are relatively constant there are three different types of effects of the treatment which could be happening (Bittner, 1981). Performance during the treatment (D) could differ from baseline performance ((B + A)/2) by an additive constant, by a multiplicative constant, or by a combination of these as shown in the upper part of Figure 11. The former type of effect indicates that all subjects were affected equally by the treatment. The latter types of effects could occur if the treatment affected the top performers more (or less) than others as illustrated in Figure 11. Analysis of covariance would be an appropriate tool to use if these latter types of effects were occurring. With the tools discussed in the preceeding paragraphs of this section, an experimenter can construct answers to many questions about his results. For instance, what was the mean effect of the treatment?

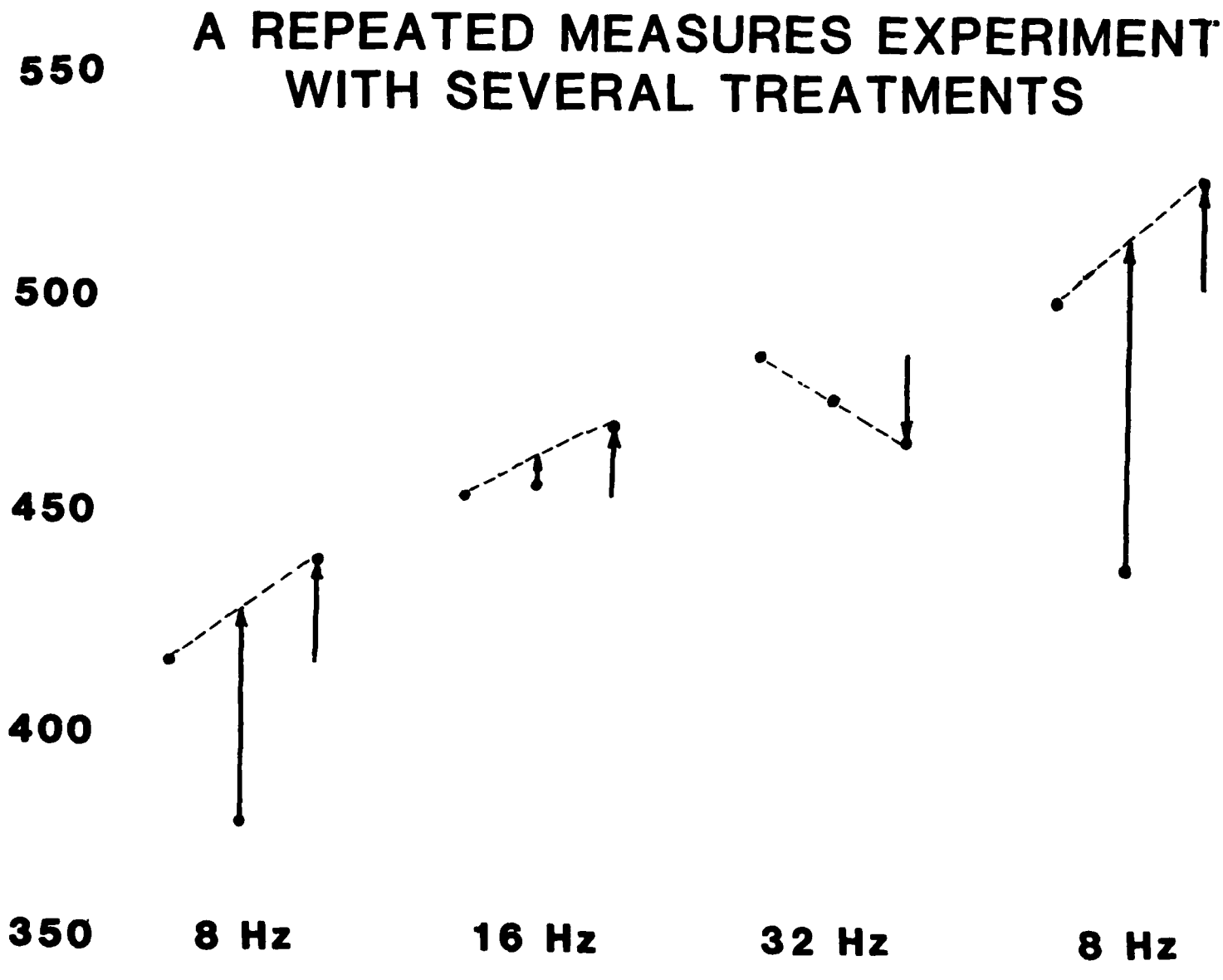


Figure 8. Example of Relative Constancy of (8 Hz (D - (B + A)/2)) Contast Across Four Experiments with an Overall Learning Trend.

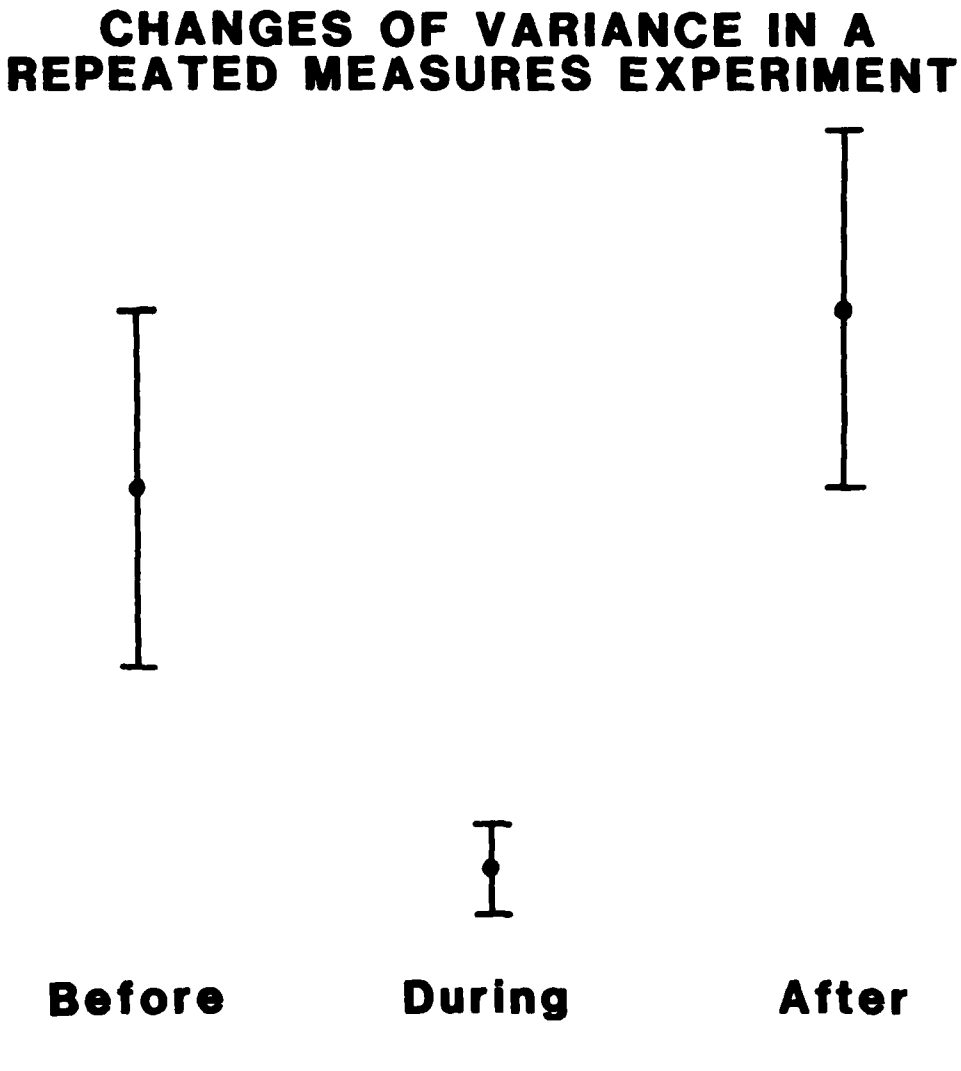


Figure 9. Illustration of Changes of Variance.

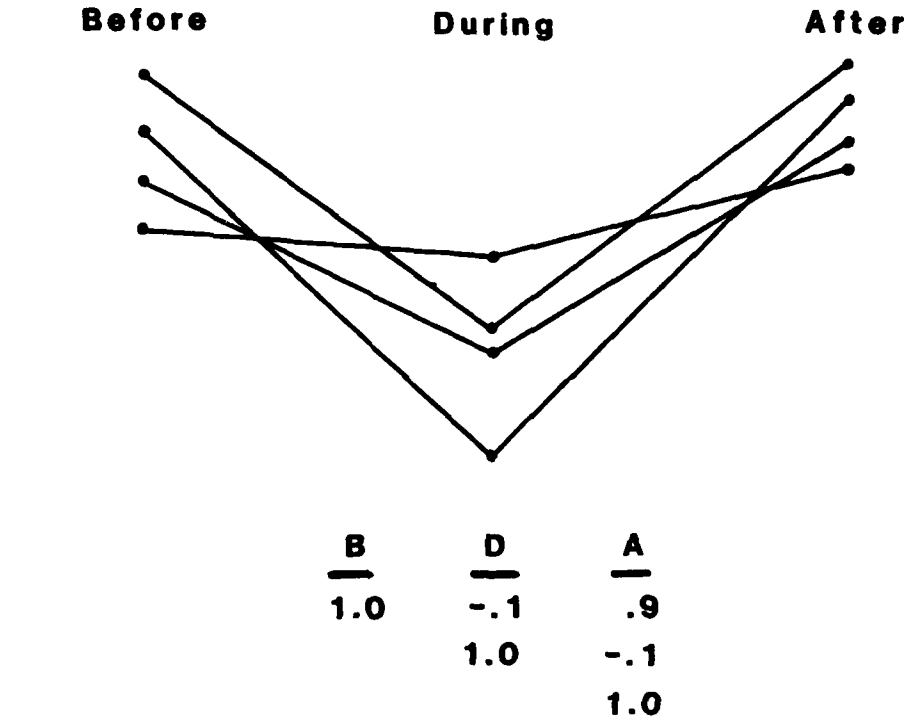


Figure 10. Illustration of Subjects Responces for Changed Intertrial Correlations.

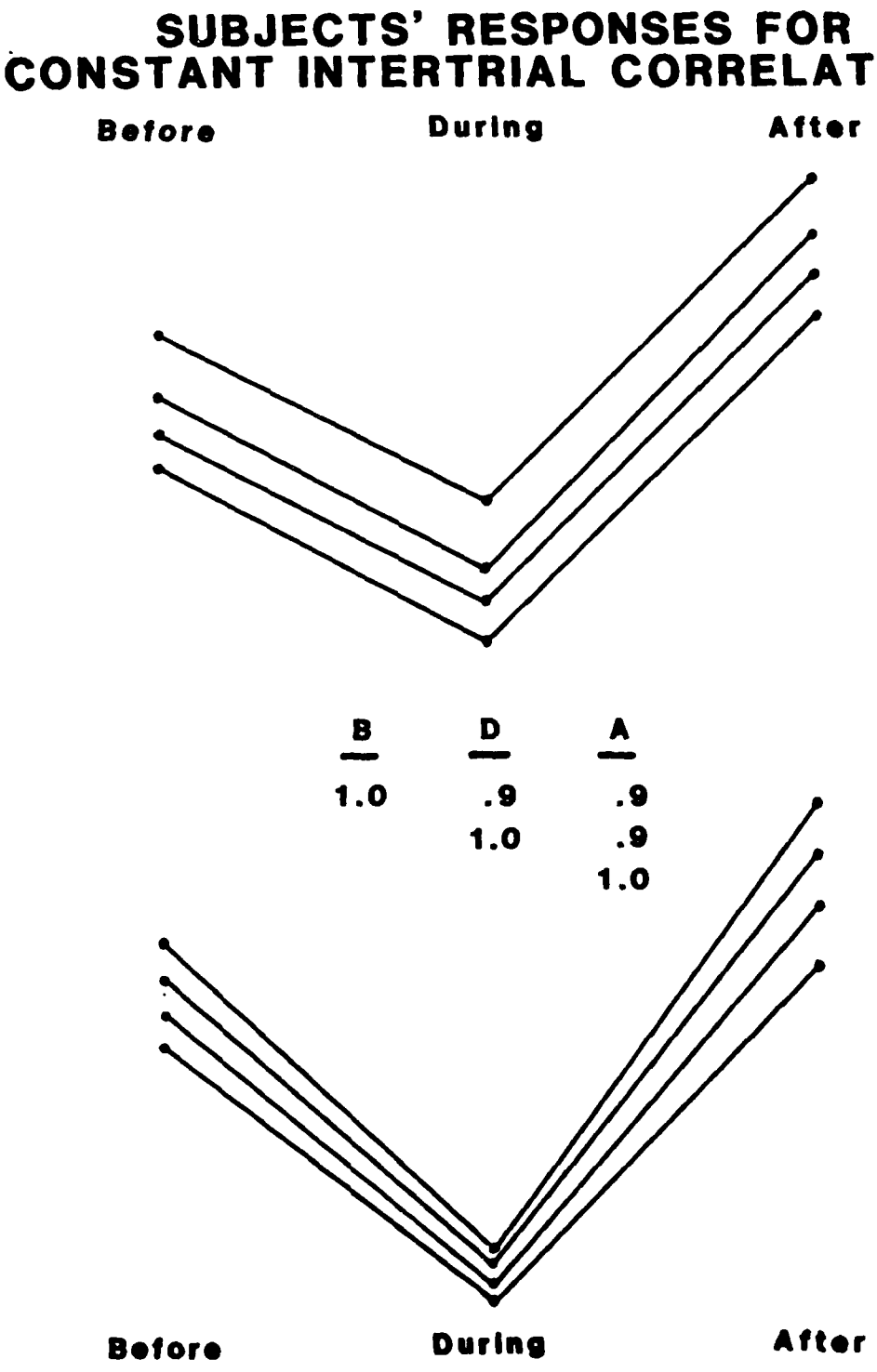


Figure 11. Illustration of Constant Intertrial Correlations with Constant and Changing Variances.

Was the effect consistant for all subjects? Was the effect proportional to baseline performance? Was the variability of performance changed by the treatment? How does the effect of one treatment compare with the effect of another? Ordinarily, these questions have not all been considered, perhaps because appropriate tools were not at hand.

Experiments with a Single Subject or Team

It is possible to obtain much valuable information from data on single subjects. If more than one subject were available, comparisons between results for each subject indicate the generalizability of the results. Tools applicable for analyzing single-subject data are presented in detail by Box and Jenkins (1970) and Glass, Wilson, and Gottman (1975). Collectively, these tools constitute an approach to analysis of "time series." They assume a series of at least 50 observations at approximately equal intervals of time representing a process which has some unchanging statistical properties described by the references cited. The criteria for evaluating the stability of means, variances, and intertrial correlations described earlier provide a basis for making the statistical assumptions of time-series analysis.

Time-series tools can be used to infer changes of mean level, slope, variance, or even more subtle characteristics of the subject's responses. Furthermore, the dynamics of the response to treatments can be studied without the loss of fidelity caused by aggregating several subjects' data. Time-series methods can also be used to study cycles of behavior, to examine feedback among several variables, or to forecast performance in the future. Generally, the time-series methods discussed in this report consist of finding a stochastic model for the data.

For example, Glass, Wilson, and Gottman (1975) offer a research tool for showing whether the variance of a time series changes in response to a treatment intervention. First a model is fit to the series, then that model is applied separately to data from before and after the intervention. The ratio of the residual variances from these two applications of the model is a statistic with an F distribution when there has been no change in variance. By comparing an empirical statistic with a table value of the F distribution, it is possible to determine whether a detectable change of variance was associated with the treatment intervention.

Time-series analysis also includes tools for investigating changes of the level of a series of observations from before to after a treatment. Time-series analysis goes beyond the usual tests of mean effects because it also characterizes how the level of the series changed over time. This branch of time-series analysis is called intervention analysis (Box & Tiao, 1975). Intervention analysis can be used for response curves deemed marginally or totally uninterpretable with traditional methods of repeated-measures analysis (Campbell & Stanley, 1966). Responses that are delayed, gradual, oscillating, or show other dynamic forms can be accommodated.

Finally, it is possible that a treatment alters the dynamics of a series of performance measurements. That is, the form of the dependency of present responses on past responses may change. To test for this eventuality, a stochastic model is fit to the observations made before the

treatment. The same model is applied to the observations made after the treatment. The auto-correlations (Box & Jenkins, 1970) $\underline{r}_i$ of the residuals from modeling the observations after the treatment are combined using Box's formula:

$$Q = T(T+2)\left[\sum_{i=1}^{k}\left(\frac{1}{T-i}\,\underline{r}_i^{\,2}\right)\right] \qquad (1)$$

where k ($\geq$ 20) is the number of autocorrelations, T is the number of observations, and p is the number of parameters in the time-series model. If the dynamics of the two series are the same, Q is chi-square distributed with k - P degrees of freedom. Hence, if Q is statistically significant, then a change of the dynamics of performance is indicated. This tool might be used, for example, to determine whether the form of neurophysiological evoked potential measurements is altered by exposure to impact.

SUMMARY OF THE BAG OF RESEARCH TOOLS

Table 3 is an inventory of the bag of research tools that has been assembled for repeated measurements. In the left hand column, an application of each tool is given. Tools are described in the right hand column. This bag of tools, as with most, is incomplete and contributions are welcome.

Table 3: A Tool Bag for Repeated Measurements

APPLICATION	TOOL
Evaluate a task's suitability for repeated measurements	Widely distributed (e.g., daily) measurements made in standard conditions (e.g., Kennedy & Bittner, 1980)
Check for stability of means (linear trend or no trend with practice)	Repeated measures ANOVA with orthogonal trend analysis (e.g., Dixon & Brown (1977) BMDP2V)
Check for stability (homogeneity) of variances	Analytic tests (e.g., Fmax) for equality of variances (Hollander, Wolfe, 1973; Scheffé, 1959; Winer, 1971)
Check for differential stability of intertrial correlations	Steiger (1980a, 1980b) CORRMAT

Table 3 Continued

Represent effects of treatments, practice within experiments, and practice between experiments	Experimental designs involving one measurement Before (B), During (D), and After (A) the treatment. Contrasts: D-(B + A)/2 represents treatment effects; B-A represents within-experiment practice; and (B' + A') - (B + A) represents practice effects between conditions. These contrasts are merely linear and quadratic trends in ANOVA. (Guignard, Bittner, & Carter, 1981)
Represent treatment effects in which the effect, D - (B + A)/2 has a non-unitary proportional component relative to (B + A)/2	Analysis of Covariance and Effect Models (Bittner, 1981; Winer, 1971)
Check for consistency of treatment effect (i.e., D - (B + A)/2 = K for all subjects?)	CORRMAT on BDA three-trial correlation matrix
Represent treatment effects in a single-subject experiment	Time-Series intervention analysis (Box & Tiao, 1975)
Test for change of variance of a single subject's performance a from before to after a treatment	Glass, Gottman, and Wilson's (1975) F-test
Test for change of a subject's response dynamics from before to after a treatment	Box's test of autocorrelation (Box & Jenkins, 1970)
Account for autocorrelations and biological cycles in repeated-measures data	Box-Jenkins (1970) stochastic time-series models

References

Alvares, K. M., & Hulin, C. L. Two explanations of temporal changes in ability-skill relationships: A literature review and theoretical analysis. Human Factors, 1972, 12, 295-308.

Bittner, A. C., Jr. Statistical tests for differential stability. Proceedings of the 23rd Annual Meeting of the Human Factors Society, Boston, MA., October, 1979, 541-545.

Bittner, A. C., Jr. Use of proportion-of-baseline measures in stress research. In G. Salvendy and M. J. Smith (Eds.), Machine Pacing and Occupational Stress. London: Taylor & Francis, 1981, 177-183.

Bittner, A. C., Jr., Lundy, N. C., Kennedy, R. S., & Harbeson, M. M. Performance Evaluation Tests for Environmental Research (PETER): Spoke tasks. Perceptual and Motor Skills, in press.

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

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

Box, G. E. P., & Tiao, G. C. Intervention analysis with applications to economic and environmental problems. Journal of the American Statistical Association, 1975, 70, 70-79.

Campbell, D. T., & Stanley, J. C. Experimental and Quasi-Experimental designs for research. Chicago: Rand McNally, 1966.

Carter, R. C. Physiological and performance measurements: A time-series model. Preprints of the 51st Annual Meeting of the Aerospace Medical Association, Anaheim, CA., May, 1980, 161-162.

Carter, R. C., Kennedy, R. S., & Bittner, A. C., Jr. Selection of performance evaluation tests for environmental research. Proceedings of the 24th Annual Meeting of the Human Factors Society, Los Angeles, CA, October, 1980, 320-324.

Carter, R. C., Kennedy, R. S., & Bittner, A. C., Jr. Grammatical reasoning: A stable performance yardstick. Human Factors, 1981, 23, 587-591.

Dixon, W. J. &, Brown, W. J., (Eds.) BMDP Biomedical Computer Programs (P-Series). Los Angeles: University of California Press, 1977.

Estes, W. K. The problem of inference from curves based on group data. Psychological Bulletin, 1956, 53, 134-140.

Fisher, R. A. The design of experiments . New York: Hafner, 1935 (1st Ed.), 1966 (8th Ed.).

Fitts, P. M. &, Posner, M. I. Human Performance, Belmont, CA: Brooks-Cole, 1967.

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

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

Guignard, J. C., Bittner, A. C., Jr., & Carter, R. C. Methodological investigation of vibration effects on performance of three tasks. Proceedings of the 25th Annual Meeting of the Human Factor Society, Rochester, N. Y., October 1981, 342-346.

Hollander, M., & Wolfe, D. A. Nonparametric statistical methods. New York: Wiley, 1973.

Jones, M. B. A two-process theory of individual differences in motor learning. Psychological Review, 1970, 77, 353-360. (a)

Jones, M. B. Rate and terminal processes in skill acquisition. American Journal of Psychology, 1970, 83, 222-236. (b)

Jones, M. B. Individual differences. In R. N. Singer (Ed.), The psychomotor domain. Philadelphia: Lea & Febiger, 1972.

Jones, M. B. Stabilization and task definition in a performance test battery. (Monograph No. NBDL-M-0001), New Orleans, LA: Naval Biodynamics Laboratory, 1980.

Jones, M. B. Differential retention and convergence with practice. First Quarterly Report on Contract No. MDA903-81-C-0293 with the Army Research Institute for the Behavioral and Social Sciences, Alexandria, VA., 27 July 1981.

Jones, M. B., Kennedy, R. S., & Bittner, A. C., Jr. A video game for performance testing. American Journal of Psychology, 1981, 94, 143-152.

Jöreskog, K. G. Statistical analysis of sets of congeneric tests. Princeton, N.J. Educational Testing Service, December, 1969 (RB-69-97).

Kennedy, R. S. A retrospective view of the "PETER" Program. Paper presented at the International Workshop on Research Methods in Human Motion and Vibration Studies. New Orleans, LA., September, 1981.

Kennedy, R. S., & Bittner, A. C., Jr. The development of a Navy Performance Evaluation Test for Environmental Research (PETER). In L.T. Pope & D. Meister (Eds.) Productivity Enhancement: Personnel Performance Assessment in Navy Systems, Navy Personnel Research and Development Center, San Diego, CA., 1977, NTIS AD A045047.

Kennedy, R. S. & Bittner, A. C., Jr. Development of Performance Evaluation Tests for Environmental Research (PETER): Complex counting test. Aviation, Space and Environmental Medicine, 1980, 51, 142-144.

Kennedy, R. S., Carter, R. C., & Bittner, Jr., A. C. A catalogue of performance evaluation tests for environmental research. Proceedings of the 24th Annual Meeting of the Human Factors Society, Los Angeles, CA., October, 1980, 344-348.

McCauley, M. E., Kennedy, R. S., & Bittner, A. C., Jr. Development of Performance Evaluation Tests for Environmental Research (PETER): time estimation test. Perceptual and Motor Skills, 1980, 51, 655-665.

Posner, M. I. Chromometric explorations of mind. Hillsdale, N.J.: Erlbaum, 1978.

Reynolds, B. The effects of learning on the predictability of psychomotor performance. Journal of Experimental Psychology, 1952, 44, 189-198.

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

Shannon, R. H. A factor analytic approach to determining stability of human performance. Proceedings of the 13th Annual Meeting of the Human Factors Association of Canada, Pt. Ideal, Ontario, September, 1980.

Shannon, R. H., Carter, R. C., & Boudreau, A. Y. A systematic approach to battery development and testing within unusual environments. Paper presented at the International Workshop on Human Motion and Vibration Studies. New Orleans, LA., September 1981.

Steiger, J. H. Tests for comparing elements of a correlation matrix. Psyhcological Bulletin, 1980, 87, 295-251. (a)

Steiger, J. H. Testing pattern hypotheses on correlation matrices: Alternative statistics and some empirical results. Multivariate Behavioral Research, 1980, 15, 335-352. (b)

Sutcliffe, J. P. On the relationship of reliability to statistical power. Psychological Bulletin, 1980, 88, 509-515.

Thorndike, R. L. Personnel selection test and measurement techniques. New York: Wiley, 1949.

Winer, B. J. Statistical principles in experimental design (2nd Ed.). New York: McGraw-Hill, 1971.


FILMED
5-8