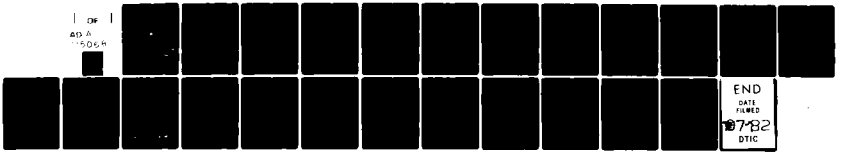
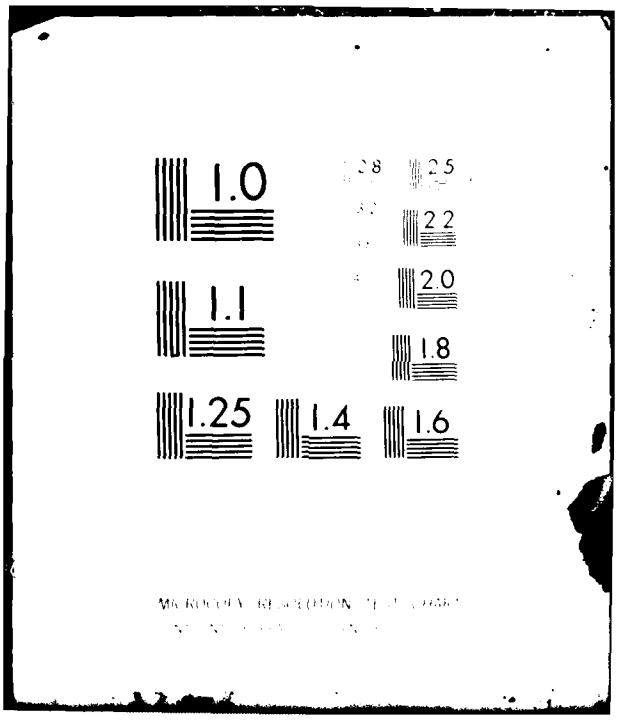
AD-A115 068 NAVAL BIODYNAMICS LAB NEW ORLEANS LA F/G 5/10
PERFORMANCE TESTS FOR REPEATED MEASURES: MORAN AND COMPUTER BAT--ETC(U)
NOV 81 A C BITTNER, R C CARTER, M KRAUSE
UNCLASSIFIED NBDL-81R012 NL




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5

(30

NBDL - 81R012

AD A115068

Performance Tests for Repeated Measures: Moran and Computer Batteries

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

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: 81R012	2. GOVT ACCESSION NO.: AD-A115 068
	3. RECIPIENT'S CATALOG NUMBER:
4. TITLE *(and Subtitle)*: Performance Tests for Repeated Measures: Moran and Computer Batteries	5. TYPE OF REPORT & PERIOD COVERED: Research Report
	6. PERFORMING ORG. REPORT NUMBER: NBDL 81R012
7. AUTHOR(s): Alvah C. Bittner, Jr., Robert C. Carter, and Michele Krause	8. CONTRACT OR GRANT NUMBER(s):
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Naval Biodynamics Laboratory, P.O. Box 29407, New Orleans, Louisiana 70189	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: MF58.524-002-5027
11. CONTROLLING OFFICE NAME AND ADDRESS: Naval Medical Research & Development Command, Bethesda, MD 20014	12. REPORT DATE: November 1981
	13. NUMBER OF PAGES: 20
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)*

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)*

Test battery, human performance tests, repeated measures, computer generated tests, factor-referenced ability tests, environmental research

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

This investigation was directed at statistical baseline evaluation of nine tasks for suitability for repeated measures applications to environmental investigations. In the first study, tasks from Moran and Mefferd (1959) were administered to 18 subjects daily for 13 work days (Monday to Friday). In the second study, Carter and Sbisa (1981) computer battery tasks were administered daily to 17 subjects (12 in common with first study) for 15 work days. Examination of the means, variances, interday reliabilities and cross-task correlations led to the recommendation of four tasks for repeated measures applications

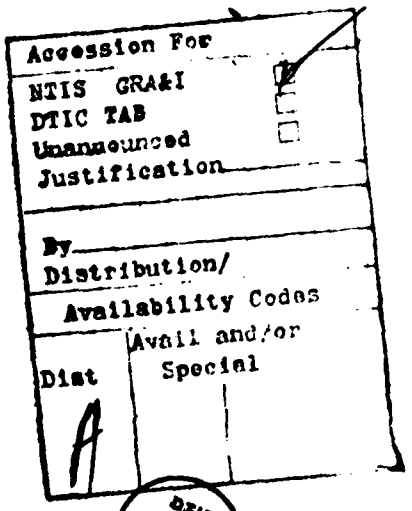
DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED

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

BLOCK 20 - ABSTRACT Continued

Vertical Addition (Nv), Perceptual Speed (PS), Grammatical Reasoning (GR), and Flexibility of Closure (FC).



NBDL - 81R012

Performance Tests for Repeated Measures: Moran and Computer Batteries

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

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

INTRODUCTION

Investigations of adverse environmental effects almost exclusively employ repeated measures of subjects (Kennedy & Bittner, 1977). The general approach in such studies is to collect data on one or more trials conducted Before, During, and After exposure. Interest in time-course effects frequently dictates a Before-During-After (BDA) experiment, but financial and statistical arguments for economy can also be made for this paradigm (Campbell & Stanley, 1963; Winer, 1971). In addition, shortages of qualified research subjects and simulator space can make experiments with independent groups infeasible. However, the most potent argument for repeated measures often rests on the requirement to minimize subject risk in hazardous environments. This argument presumes that with tasks which are suitably stable for repeated measurements the number of trials and thus, subject exposure, will be reduced because individual differences may be removed from the error variance. Clearly, statistical suitability should be sought before a task is used in a BDA experiment where risk minimization is a serious consideration. A program to evaluate the suitability of performance tests for repeated measures and to develop methodologies for test applications has been underway for the last four years (Kennedy, & Bittner, 1977; Carter, Kennedy, & Bittner, 1980; Kennedy, Carter, & Bittner, 1980; Bittner, 1981a; Bittner & Carter, 1981; Guignard, Bittner, & Carter, 1981). This investigation is directed at the baseline evaluation of tasks drawn from the Moran Battery (Moran & Mefferd, 1959) and the Carter and Sbisa Computer Generated Battery (1981) as part of this program.

According to Jones (1972, 1980), candidate tests for repeated measures studies should meet rigorous statistical qualifications. Meaningful repeated measurements generally require that the means variances, and intertrial correlations of a test be well-behaved when obtained under constant conditions. Specific statistical characteristics considered necessary are as follows: (1) the means change in a linear manner or are unchanging over trials; (2) variances are homogeneous over trials; and (3) intertrial correlations are differentially stable. Pertinently, the criteria for the means follows the considerations of Campbell and Stanley (1963, p. 38), and those for variances and correlations are implied by traditional assumptions for related measures analysis of variance (e.g., Scheffé, 1959; Winer, 1971). Constancy of correlation, in addition to being a traditional assumption, assures that the same attribute is being measured on each occasion of measurement (Alvares & Hulin, 1972). Without such constancy, attribution of effect and scientific generalization may be impossible (Bittner, 1979; Jones, Kennedy, & Bittner, 1981). The present investigation was directed at determining when, if ever, in practice tasks obtain the desired statistical characteristics under baseline conditions.

---

[1] This work was funded by the Naval Medical Research and Development Command and was performed under Navy Work Unit No. MF58.524.002-5027. The research program was identified as the Performance Evaluation Tests for Environmental Research (PETER) Program in earlier reports. The opinions are those of the authors, and do not necessarily reflect those of the Department of the Navy. Requests for reprints may be sent to Dr. Alvah C. Bittner, Jr., Naval Biodynamics Laboratory, Box 29407, New Orleans, LA 70189.

Several investigations are relevant to the present study. Horn (1972), has reported factor analyses which included three cognitive tasks from the Moran battery in a monumental investigation of 16 tasks administered daily for 10 days to 106 subjects. These analyses were designed to separate "trait" from "state" contributions to task variation. Horn identified Perceptual Speed, Visualization, and Flexibility of Closure as having state variation. However, state contributions were typically smaller than trait contributions. Unfortunately, the Horn (1972) results were not reported in sufficient detail to permit mean, variance, and differential stability analyses of the types identified above. More recent investigations in this laboratory have reported results of stability analyses for tasks similar to those included in the Computer Battery (Seales, Kennedy, & Bittner, 1980; Carter, Kennedy, & Bittner, 1981). Seales, et al. (1980) reported that a 10 minute arithmetic test (composed of successive addition, subtraction, multiplication, and division subtasks) possessed mean, variance, and differential stability from the first day of a 15 day study. The average correlation across days was $\bar{r} = 0.94$ for the total correct score. More recently Carter, et al. (1981) have reported that the Grammatical Reasoning Task (Baddeley, 1968) met all stability criteria after only four daily (60 second) administrations; with a reliability of $\bar{r} = 0.82$ across the differentially stabilized trials. The Horn (1972), Seales, et al. (1980), and Carter, et al. (1981) results encouraged the present multi-task investigation.

One purpose of this investigaton was to evaluate the statistical characteristics of tasks drawn from the Moran (Moran & Mefferd, 1959) and Computer Batteries (Carter & Sbisa, 1981). A second purpose was to explore the relationships between tasks subsequent to their becoming differentially stable.

METHOD

The approach in this investigation was to conduct two sequential experiments with each directed at a specific battery. In the first experiment, tasks from the Moran Battery were studied and, in the second, tasks from the Computer Battery were investigated. The two experiments are described sequentially in the following sections.

Experiment 1: Moran Battery

Tasks

The Moran Battery employed in this study consisted of five simple paper-and-pencil tests which were constructed to follow the format in French's (1954) kit of reference aptitude and achievement factors (Moran & Mefford, 1959; Moran, Kimble, & Mefferd, 1964). Twenty alternate forms for each task with accompanying instruction sheets and practice problems are available. The tasks measured included: Flexibility of Closure (FC), Number Facility (NF), Perceptual Speed (PS), Speed of Closure (SC), and Visualization (V). Copies of the alternate forms were obtained from Moran.

> Flexibility of Closure (FC). This task required retaining the image of a specified configuration despite the influence of other distracting configurations in the perceptual field (Moran & Mefferd, 1959). The specific configuration was given in this

task, unlike the situation for Speed of Closure, in the form of 36 geometric figures to be copied onto matrices of dots. Ekstrom French, Harmon, and Dermen (1976) placed this task under a general Closure, Flexibility of (CF) factor together with Hidden Figures and Hidden Patterns tasks. In addition, they gave a brief review and 26 references in a format followed for all their referenced factors. The FC score was the number of figures correctly copied in 180 seconds.

Number Facility (NF). This test required the addition of one or two digit numbers in sets of three (Moran & Mefferd, 1959). Ekstrom, et al. (1976) placed a similar task under a general Number Facility (N) factor together with Division, Substraction, and Multiplication, and Addition and Subtraction Correction Tests. The NF test score was number of correct answers in 180 seconds.

Perceptual Speed (PS). This task required the crossing out of every digit that was like one circled at the beginning of that row in a row of 30 digits (Moran & Mefferd, 1959). It appears to fall under the general Ekstrom, et al. (1976) Perceptual Speed (P) factor which was identified by Finding A's, Number Comparison, and Identical Picture Tests. Ekstrom, et al. also provided 70 references to studies identifying Perceptual Speed. The score of the Moran and Mefferd task was the number of digits correctly marked in 150 seconds.

Speed of Closure (SC). This task required the search for simple four-letter words imbedded in fields of random letters which did not form unintended words (Moran & Mefferd, 1959). Words were mainly nouns, with proper names, foreign, and plural words excluded. As opposed to the Flexibility of Closure task described earlier, fore knowledge of the material to be searched was not given. Ekstrom, et al. (1976) placed this task in a general Closure, Verbal (CV) factor which was identified by Scrambled Words, Hidden Words, and Incomplete Words. The SC score was number of words correctly circled in 150 seconds.

Vizualization (V). This task required the visual following of the path of a line, from left to right, and placing the line numbers in the appropriate cell on the right (Moran & Mefferd, 1959). Sets of 10 "tangled lines" constituted the stimulus material. Comparison of this task with the factor reference cognitive tests, identified by Ekstrom, et al. (1976), suggested that this test was more related to their Spatial Scanning (SS) Factor. Maze Tracing Speed, Choosing a Path, and Map Planning tests identified the Ekstrom, et al. SS Factor which was defined as "Speed in exploring visually a wide or complicated spatial field". Scoring on the Vizualization (V) Test was the number of cells correctly numbered in 180 seconds.

Subjects

The subjects employed in this experiment were 18 volunteers from a population of enlisted men, ages 19 to 24, assigned to this laboratory as full-time research subjects. All volunteers were recruited, evaluated, and employed in accordance with procedures specified in Secretary of the Navy Instruction 3900.39 Series and Bureau of Medicine and Surgery Instruc-

tion 3900.6 Series. These instructions are based on voluntary informed consent and meet provisions of prevailing national and international guidelines. Volunteers were given cardiovascular, pulmonary, skeletal, and other examinations to insure their capability to serve in possibly hazardous environmental research; however, they were generally representative of the enlisted population in intelligence. A description of the volunteer qualification procedures appears in Thomas, Majewski, Ewing, and Gilbert (1978). described above. They had some exposure to psychological testing, mostly psychomotor, but had no previous exposure to the Moran test battery.

Procedure

Prior to practice and testing, subjects were briefed on the tasks in the experiment. A familarization practice trial was given on the following day on all tasks. Responses were checked to insure task understanding. Formal testing was then conducted for 13 work days (Monday thru Friday) with one trial per day on each task given in the order PS, SC, NF, FC, and V. Trials were conducted on separate days to avoid inflation of correlations by within day autocorrelative effects (Thorndike, 1949).

Experiment 2: Computer Generated Battery

Tasks

The computer battery employed in this study consisted of four paper-and-pencil tasks with their items randomly sampled by computer from among all items of their type (Carter & Sbisa, 1981). Because of the method of generation, a very large number of alternate forms ($>10^{35}$) may be produced along with instruction sheets, practice problems, and answer sheets. Tasks included in this study were Vertical Addition (Nv), Horizontal Addition (Nh), Number Comparison (Nc) and Grammatical Reasoning (GR).

Vertical Addition (Nv). This task required the addition of three two-digit numbers arrayed vertically. Conceptually, Nv was based on the (vertical) Addition Test described by Ekstrom, et al. (1976); however, they used three one- or two-digit numbers. The conceptual basis of this task implied that it would fall under their Number Facility (N) Factor. The Nv score was the number of correct responses during two consecutive 120 second administrations.

Horizontal Addition (Nh). This task required the addition of three three-digit numbers, ranging from 100 through 999, arranged horizontally. It was suggested by a task employed by Alluisi (1969, pp. 68-69) and was employed to see if the format altered the differential properties of the test. It was suspected that a substantial portion of this task would fall under Number Facility (N) as described by Ekstrom, et al. (1976). The Nh score was the number of correct responses during two consecutive 120 second administrations.

Number Comparison (Nc). This task required the comparison of two, 3 to 9 digit, horizontally arranged numbers and a response of S (Same) or D (Different). Modeled after the Number Comparison Test given in Ekstrom, et al., (1976), it would be expected to fall under their Perceptual Speed (P) Factor. The

Nc score was the number of correct minus number of errors for 180 seconds administration.

Grammatical Reasoning (GR). This task was based on Baddeley's three minute reasoning test (1968) and required the comparison of a statement on the order of two letters (A and B) with a displayed order. For example, "A follows B: BA" or "A is not preceeded by B: BA". The T (True) or F (False) responses were required. Thirty-two items constitute this test using affirmative or negative phrasing, active or passive voice, A or B mentioned first, the verbs "precedes" or "follows" and validity (T of F) of the comparisons. Different random item orders constituted different forms of this task. Grammatical Reasoning appeared to be related to the Ekstrom, et al. (1976) Reasoning, Logical (RL) Factor. The RL factor is defined as "the ability to reason from premise to conclusion, or to evaluate the correctness of a conclusion". The GR score was the number of correct minus wrong responses made in 90 seconds.

Subjects and Procedure

The subjects employed for this experiment were 17 volunteers from the general population described in Experiment 1. Of these subjects, 12 had previously been tested in Experiment 1. All subjects had previous psychological testing exposure, primarily psychomotor. Subjects, subsequent to briefing, were tested for 15 work days on the battery tasks administered in random order.

RESULTS

The results of the two experiments were analyzed in three phases. In the first two phases, the Moran and Computer Batteries were individually analysed. The final phase of the analysis explored the differential relationships between the two batteries.

Experiment 1: Moran Battery

The analysis of tasks was conducted in two stages focused sequentially on: (1) task differential stabilities and cross correlations; and (2) stability of means (linearity) and variances (homogeneity) over days. These will be taken up in turn, with tasks considered in the order: FC, NF, PS, SC, and V.

Differential Stability and Cross Correlations

Determination of the point in practice at which each task became differentially stable was accomplished using the methodology and general computer program developed by Steiger (1980a, 1980b). As a first step for each task, the constancy of the reliabilities over Days 1 - 13 was assessed. Failing a clearly nonsignificant ($p > .10$) result, a second analysis was conducted over Days 2 - 13 and significance was evaluated. Successive analyses were continued, dropping leading days, until indications of differential stability were obtained. Task analyses are given below and cross correlations of stabilized tasks are given subsequently.

Flexibility of Closure (FC). The differential stability test across all days (1 - 13) was very highly significant ($\chi^2(77) = 121.1$; $p < .0012$), indicating changing correlations. However, after dropping the first two days, the test yielded nonsignificant results with $\chi^2(54) = 65.8$ ($p > .13$). The estimate of the FC differentially stable reliability across Days 3 - 13 was $r = 0.882$.

Number Facility (NF). Across all days (1 - 13), the test statistic indicated changing cross day reliabilities with $\chi^2(77) = 201.4$ ($p < 10^{-9}$). The Days 9 - 13 statistic, however, was nonsignificant with $\chi^2(9) = 14.5$ ($p > .10$). The estimate of the NF differentially stable reliability across Days 9 - 13 was $r = 0.830$.

Perceptual Speed (PS). The stability test across Days 1 - 13 yielded very highly significant results ($\chi^2(77) = 135.6$; $p < .0001$). However, after dropping the first six days, the results were clearly nonsignificant ($\chi^2(20) = 23.3$; $p > .27$). The PS estimated differentially stable reliability across Days 7 - 13 was $r = 0.837$.

Speed of Closure (SC). Across all days (1 - 13), the test statistic indicated changing cross day reliabilities ($\chi^2(77) = 97.8$; $p = 0.06$). The test statistic after dropping the first day, however, was clearly nonsignificant with $\chi^2(65) = 97.8$; $p > .19$. The Days 2 - 13 SC differentially stable reliability was estimated as $r = 0.767$.

Visualization (V). The differential stability test across Days 1 - 13 yielded significant results with $\chi^2(77) = 104.0$ ($p < .022$). However, after dropping the first 5 days, the test statistic was nonsignificant ($\chi^2(27) = 36.7$; $p > .10$). The Days 6 - 13 differentially stable reliability was estimated as $r = 0.664$.

Table 1 gives estimates of cross-task correlations over respective differentially stable days as identified above. Paralleling the cross day reliability estimates, task correlations were averaged over all pairs of stable task days, excepting pairs measured on the same day. Within-day correlations were not included so as to avoid inflation with within-day state covariations shown for some of the Moran Battery tasks by Horn (1972, p. 178). The estimated NF-V correlation of $r = 0.118$, for example, was the Fisher-z average of the 35 cross correlations between their respective stable Days 9 - 13 and 6 - 13. Pertinently, the averaged cross-day correlation estimates of either differentially stable reliabilities or cross-task correlations have been shown to be markedly less variable than single estimates (Bittner, 1981b). Table 1 also summarizes the estimates of task reliabilities and provides corrected-for-attenuation estimates of cross correlations.

Analyses of Means and Variances

Analysis of the points at which means and variances became stable was accomplished respectively by: BMDP2V (Dixon & Brown, 1977); and Fmax (Winer, 1971) and regression statistical tests. Figure 1 shows means over days and was a guide for the analyses. Examining this figure, it can be seen that, generally, performance improved with practice on all tasks, indicating learning.

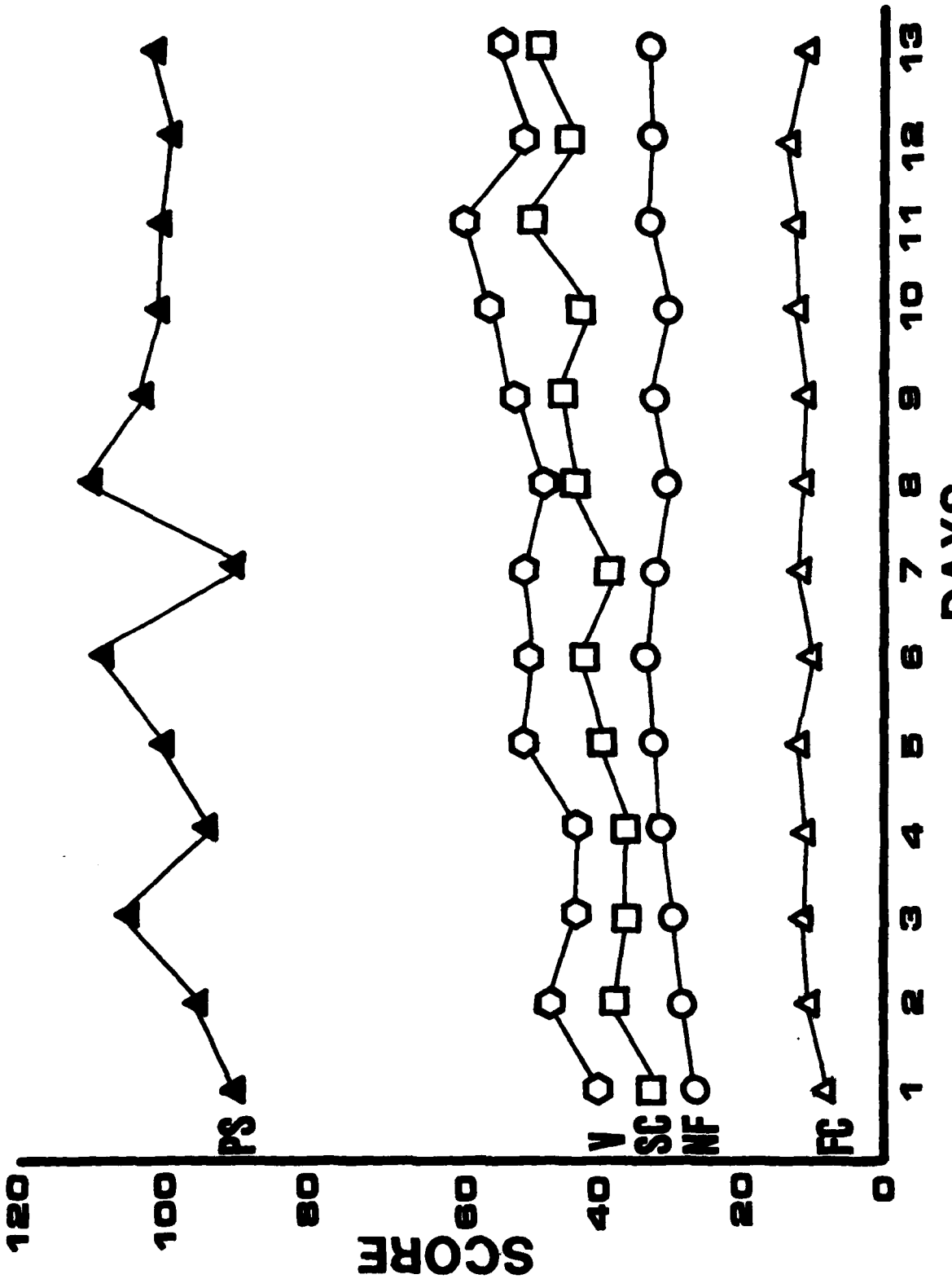


FIGURE 1. MORAN BATTERY MEANS OVER 13 DAYS (N=18)

Table 1. Moran Battery Differentially Stabilized Cross-Day Correlations*

Test	1	2	3	4	5
1) Flexibility of Closure	(0.882)	0.124	-.176	-.050	0.353
2) Number Facility	0.145	(0.830)	0.435	0.567	0.118
3) Perceptual Speed	-.201	0.511	(0.873)	0.603	0.028
4) Speed of Closure	-.061	0.711	0.737	(0.767)	-.050
5) Visualization	0.461	0.159	0.037	-.070	(0.664)

*Correlations above, reliablities along, and corrected-for-attenuation estimates below diagonal.

Flexibility of Closure(FC). Examining the FC plot, it can be seen that performance appears to improve most rapidly over the first two days and to be essentially linear thereafter. ANOVA over all days (1 - 13) was significant ($F(12,204) = 6.93$, $p < 10^{-8}$) with significant nonlinear trends ($F(11,204) = 2.72$, $p < .003$). After dropping the first two days, the significant overall effect ($F(10,170) = 11.04$; $p < 10^{-8}$) was dominated by the linear component ($F(1,17) = 14.40$; $p < .001$) with all nonlinear components nonsignificant ($p > .05$). The linear component accounted for 54.7% of the Days 3 - 13 trend. The variances were homogeneous over all days with Fmax (13,17) = 2.13 ($p > .10$); the standard-deviation across days was estimated to be 5.92. Overall, means and variances were stable from Day 3 onward.

Number Facility (NF). Figure 1 shows that NF performance generally increases with practice and appears unchanging after Day 8. ANOVA over all days (1 - 13) yielded $F(12,204) = 6.36$ ($p < 10^{-7}$) with significant nonlinear trends $F(11,204) = 3.05$; $p < .001$. After dropping the first 8 days, the overall $F(4,68) = 1.14$ was nonsignificant ($p > .34$). Although the variances were nonsignificantly heterogeneous across all days (Fmax(13,17) = 3.12; $p > .05$), it was observed that the values for the first two days ranked respectively lowest and next lowest with subsequent days appearing near asymptotic. The correlation between logarithmic (log) transformed variances and test-day number was $r = 0.538$ ($p < .06$). Dropping the first two days, the correlation dropped to 0.132 ($p > .69$) and the standard deviation was conservatively estimated as 10.35. The means and standard deviations were jointly stable after Day 8.

Perceptual Speed (PS). Figure 1 shows that PS mean, in addition to a slight overall increase, appeared to fluctuate over the course of the first eight days. ANOVA over all days (1 - 13) yielded $F(12,204) = 9.60$; $p < 10^{-8}$ with significant nonlinear trends ($F(11,204) = 10.37$; $p < 10^{-8}$) accounting for 82% of the variations. After dropping the first eight days, the means became stable and level ($F(4,68) = 0.57$; $p > .68$) with a mean of 101.7. The variances over all days, although appearing somewhat unstable, yielded a nonsignificant Fmax (13,17) = 4.34; ($p > .05$) and a nonsignificant trend test when log transformed variances were correlated with day number ($r = 0.444$; $p = .128$). Overall, means and variances were jointly stable only over Days 9 - 13.

Speed of Closure (SC). Figure 1 gives the SC means which generally show increasing performance with "cyclic" nonlinear trends. ANOVA over all days (1 - 13) yielded $F(12,204) = 22.40$ ($p < 10^{-8}$) with the nonlinear trend components significant ($F(11,204 = 5.98$; $p < 10^{-6}$). Only after dropping the first 10 days do the means appear stable and level ($F(2,204) = 2.00$; $p > .15$). The Days 11 - 13 mean was 44.85. The variances across all days (1 - 13) were homogenous (Fmax(13,17) = 2.14; $p > .10$) with an estimated standard deviation of 7.92. Jointly, the means and variances were apparently stable only across Days 11 - 13.

Visualization (V). Figure 1 shows that V performance increased over days with apparently higher order nonlinear trends. ANOVA over all (1 - 13) Days was significant ($F(12,204) = 14.80$); $p < 10^{-8}$) with significant nonlinear trends ($F(11,204) = 6.30$; $p < 10^{-9}$). The nonlinear trends continued in the means even over the last three days where the nonlinear component was still significant ($F(1,204) = 9.71$; $p < .003$). The standard deviations across all days (1 - 13) appeared to be negatively related to day number with the largest (10.00) on Day 1 and the smallest (6.12) on Day 13. This trend was confirmed by the significant ($p < .006$) correlation, $r = 0.710$, between day and log transformed variance. This variance trend appeared to continue, although not significant ($p > .05$), over the last three days $r = -.532$. Hence, conservatively, neither V mean nor variance appeared stable even over the last three days.

Experiment 2: Computer Battery

Analysis of tasks was conducted in three stages dealing sequentially with correlation, variance, and mean stability as in the first experiment. Tasks were considered in the order: Nv, Nh, Nc and GR.

Differential Stability and Cross Correlations

Determination of the point in practice at which that each task became differentially stable was accomplished using the Steiger (1980a, 1980b) based methodology employed for the first experiment. Task analyses are described below and cross correlations of stabilized tasks are given subsequently.

Vertical Addition (Nv). The Steiger differential stablity test across all (1-15) days indicated changing reliabilities $\chi^2(104) = 126.5$ ($p < .07$). However, after dropping the first two days, the test statistic became nonsignificant ($\chi^2(77) = 90.8$; $p > .13$). The Days 3 - 15 Nv differentially stable reliability was estimated as $r = 0.921$.

Horizontal Addition (Nh). Indicating stability from the first day, the Nh test statistic across all days was nonsignificant with $\chi^2(104) = 105.4$ ($p > .44$). The Days 1 - 15 Nh differentially stable reliability was estimated as $r = 0.785$.

Number Comparison (NC). The NC Steiger analysis across all days (1 - 15) was very highly significant ($\chi^2(104) = 142.5$; $p < .008$). However, after dropping the first two days, the test statistic became nonsignificant ($\chi^2(77) = 88.8$; $p > .16$). The Days 3 - 15 NC differentially stable reliability estimate was $r = 0.766$.

Grammatical Reasoning (GR). The GR test statistic over all days (1 - 15) was found significant ($\chi^2(104) = 129.7$; $p < .05$). After dropping the first four days, the statistic became nonsignificant with $\chi^2(54) = 64.4$ ($p > .15$). The Days 6 - 15 differentially stable reliability estimate was $r = 0.874$.

Table 2 gives estimates of task cross correlations over respective differentially stable days as identified above. Table 2 also summarizes the estimates of task reliabilities and provides corrected-for-attenuation estimates of cross correlations.

Analyses of Means and Variances

Figure 2 shows mean performances over days and was a guide for the analyses. Examining this figure, it can be seen that, generally, performance increased with practice on all tasks which indicated learning. In the following Vertical Addition (Nv), Horizontal Addition (Nh), Number Comparison (NC) and Grammatical Reasoning (GR) will be considered in turn.

Vertical Addition (Nv). Mean Nv performance, with the exceptions of irregularities at Days 4 and Days 11 - 12, appears linear subsequent to Day 2. ANOVA over all days (1 - 15) reveals a significant effect $F(14,224) = 9.12$ ($p < 10^{-9}$) with both linear ($F(1,16) = 23.79$; $p < .0002$) and nonlinear components ($F(13,224) = 3.60$; $p < .0001$) clearly significant. After dropping Days 1-4, ANOVA over the remaining days (5 - 15) was significant ($F(10,160) = 5.29$; $p < 10^{-6}$) with the linear component significant, explaining 59% of the variance, and the nonlinear component also significant ($F(9,160) = 2.40$; $p < .02$). Significant nonlinear components manifested themselves until after dropping Days 1-11 where the overall ANOVA was nonsignificant ($F(3,48) = 2.31$; $p > .088$). Hence, level and stable means are indicated only after dropping Days 1 - 11. The omnibus $F\max(15,16) = 2.48$; $p > .1$) was nonsignificant; however, the correlation between log variance and day was $r = 0.894$ ($p < 10^{-5}$). After dropping Days 1-8, this correlation dropped to 0.442 and after dropping Days 1 - 9, the correlation was 0.274 ($p > .599$). The Days 10-15 estimated standard deviation was 8.00. Altogether Nv means and variances were jointly stable after Day 11.

Horizontal Addition (Nh). Figure 2 shows Nh mean performance increasing over the first three days with apparently level performance thereafter. ANOVA across all days (1 - 15) yielded a significant effect $F(14,224) = 5.94$; $p < 10^{-7}$) with significant nonlinear trends ($F(13,224) = 2.40$; $p < .005$). Dropping the first three days, the overall $F(11,176) = 1.04$ was clearly nonsignificant ($p > 0.41$), supporting the view of level performance over Days 4 - 15. The omnibus $F\max(15,16) = 3.17$ was nonsignificant ($p > .05$), but a

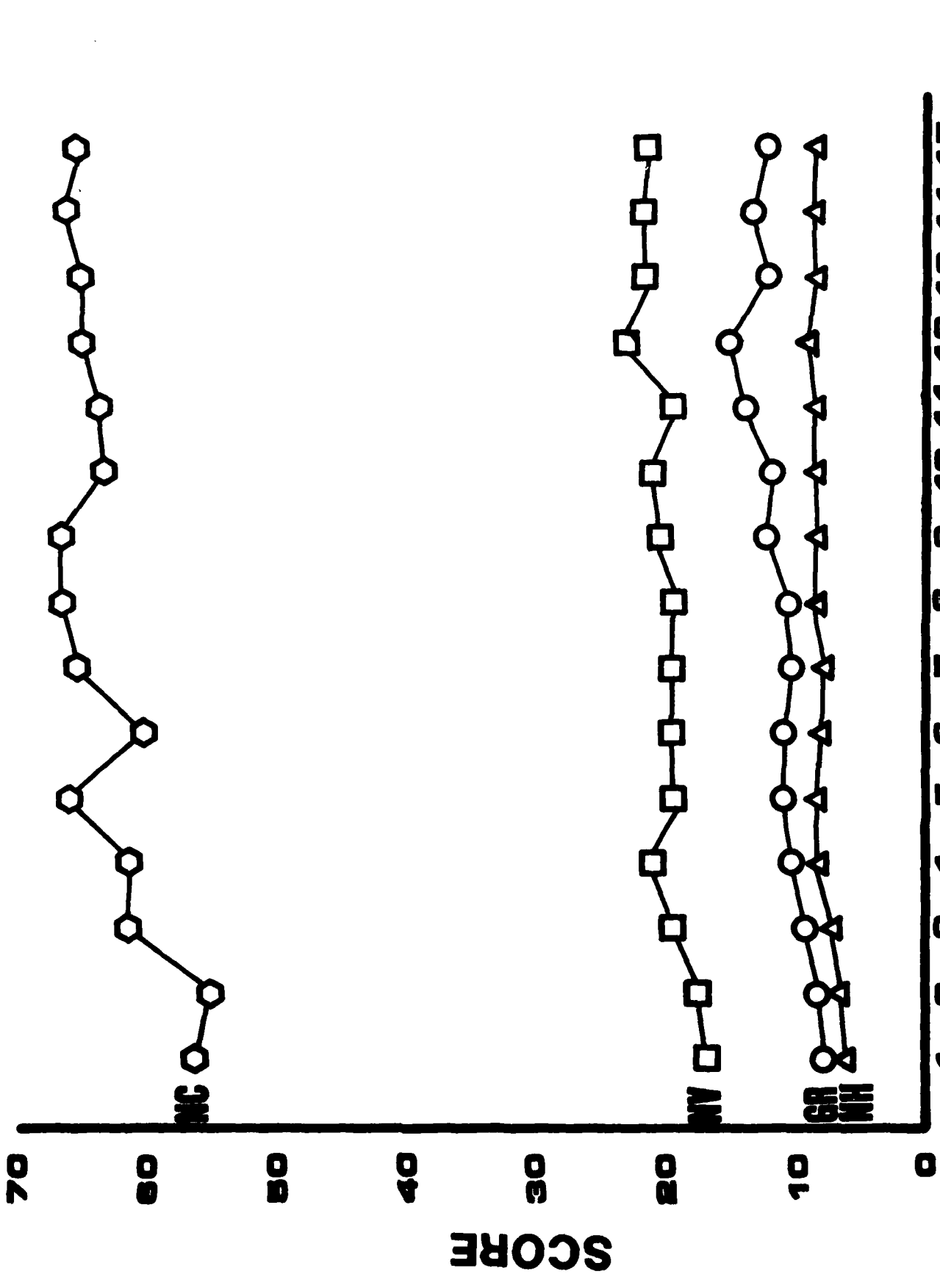


FIGURE 2. COMPUTER BATTERY MEANS OVER 15 DAYS (N=17)

Table 2. Computer Battery Differentially Stabilized Cross Day Correlations*

Test	1	2	3	4
1) Vertical Addition	(0.921)	0.790	0.606	0.209
2) Horizontal Addition	0.929	(0.785)	0.643	0.261
3) Number Comparison	0.721	0.829	(0.766)	0.259
4) Grammatical Reasoning	0.232	0.315	0.317	(0.874)

*Correlations above, reliabilities along, and corrected-for-attenuation estimates below diagonal.

trend of increasing variances with days was apparent. The correlation of log transformed variances and days was $r = 0.867$ ($p < 10^{-4}$). This trend for correlated log variances and days continued over Days 13 - 15 with $r = 0.999$, ($p < .03$). Altogether, the Nh means are stable after three days, but variances appear to be increasing across all days.

Number Comparison (NC). Figure 2 shows NC performance increasing nonlinearly over the first 6 days and maintaining a level thereafter. ANOVA over all days (1 - 15) yielded a significant $F(14,224) = 3.88$ ($p < 10^{-5}$) with significant $F(13,224) = 2.00$ ($p < .022$) nonlinear trends. However, after dropping the first 6 days, the overall $F(8,128) = 0.51$ was clearly nonsignificant ($p > .84$) and confirmed the graphical impression. Although the variances were nonsignificantly heterogeneous across days by the omnibus Fmax (15,16) = 4.21 ($p > .05$), the correlations of log transformed variance and day number was $r = 0.777$ ($p < .0007$). This trend of the variances appeared present until after dropping Days 1 - 9 where $r = 0.136$ was nonsignificant ($p > .79$) with an estimated standard deviation of 16.65. Altogether NC means and variances were jointly stable after Day 9.

Grammatical Reasoning (GR). Mean performance for GR appears to increase with negative acceleration across all days (1 - 15). Confirming this view, ANOVA across all days revealed significant linear ($F(1,16) = 20.49$; $p < .0005$) and quadratic ($F(1,16) = 4.88$; $p < .045$) trends. Over Days 2 - 15, the overall ANOVA was significant ($F(13,208) = 3.80$; $p < 10^{-4}$). The linear component

Table 3: Moran and Computer Battery Differentially Stabilized Correlations

Variable		V	FC	SC	NF	PS	NC	GR	VA	HA
		1	2	3	4	5	6	7	8	9
V	1	1.000								
FC	2	.230	1.000							
SC	3	- .111	- .224	1.000						
NF	4	.061	.073	.497	1.000					
PS	5	.025	- .317	.548	.304	1.000				
NC	6	.204	.008	.342	.541	.315	1.000			
GR	7	- .501	.023	.055	.199	- .255	.229	1.000		
VA	8	- .011	- .206	.565	.840	.426	.597	.201	1.000	
HA	9	- .045	- .095	.396	.704	.302	.635	.253	.782	1.000

Moran and Computer Batteries
13

accounted for 68.2% of the sum of squares ($F(1,16) = 15.47$ ($p < .001$), with nonsignificant nonlinear trends ($F(12,208) = 1.31$; $p > .21$). The omnibus $Fmax(15,16) = 4.05$ ($p > .05$) was nonsignificant; however, the correlation between log variance and day was $r = 0.541$ ($p < .037$). After dropping the first two days, the correlation was $r = 0.185$ which was nonsignificant ($p > .54$). Hence, over Days 3 - 15 both GR means and variances were stable.

Moran and Computer Battery Differential Relationship

Table 3 gives estimates of cross-task correlations over differentially stable days identified in the analyses of the individual batteries. Based upon 12 subjects common to both the Computer and Moran Battery studies, the pattern and range of values ($r = -0.50$ to 0.84) suggested several common factors. The factor structure was explored by factor analysis.

An iterated principal-factor analysis (PFA) was performed on the correlation matrix of the two batteries using BMDP4M (Dixon & Brown, 1977). Subsequent to identification of three factors with eigen values greater than unity by Principal Components Analysis, PFAs were run sequentially using the commonality estimates resulting from each analysis as input to the next analysis. The sequence of PFAs was continued until the maximum commonality change was less than 0.005. This procedure, it is noteworthy, yields results equivalent to those obtained by MINRES Analysis (Harmon, 1976). Table 4, in addition to commonality, gives the factor loadings subsequent to Varimax Rotation.

Table 4 shows that Factor 1 (F1) explains more than twice the variance of Factors 2 and 3 (F2 and F3). Explaining 34.6% of the possible variation, F1 is dominated by loadings of 0.889 for Nv, 0.858 for NF, 0.819 for Nh, and 0.712 for NC. The heavy loadings for the three arithmetic tasks indicate that F1 is related to the Egstrom et al. (1976) Number Facility (N) factor and support naming it "Number Facility". Factor 2 (F2) explains 14.8% of the

Table 4. Rotated Factor Loadings and Commonalities

Variable		Factor 1	Factor 2	Factor 3	Commonalities
V	1	.148	.878	-.289	.8764
FC	2	.016	.120	-.473	.2381
SC	3	.471	-.035	.578	.4811
NF	4	.858	-.022	.076	.7428
PS	5	.303	.240	.753	.7161
NC	6	.712	.062	.028	.5109
GR	7	.269	-.660	-.222	.5843
VA	8	.889	-.076	.296	.8839
HA	9	.819	-.128	.119	.7015
VR		3.112	1.333	1.300	

possible variation. A bipolar factor, F2 is identified by loadings of 0.878 for V and -0.680 for GR. The opposition of spatial and verbal tasks suggests that this may represent a cognitive style factor which might be named "Reasoning vs Visualization". The last factor (F3) accounted for 14.1% of the possible variation and is identified by two positive loadings with 0.753 for PS and 0.518 for SC. The failure of NC to appear with PS on this factor contraindicates the association of this variable with the Egstrom, et al. (1976) Perceptual Speed (P) factor which was identified by both PS and NC. F3 will be named "Perceptual Speed Task". Altogether, the three factors explain 63.8% of the possible obtainable variance.

DISCUSSION

This investigation was directed at the evaluation of tasks for repeated measures application to environmental investigations. Drawn from the Moran (Moran & Mefferd, 1959) and Computer (Carter & Sbisa, 1981) Batteries, nine tasks were examined with respect to the points in practice at which that they obtained unchanging or linearly changing means, homogeneous variances, and constant (differentially stable) intertrial correlations. The relationships between tasks, subsequent to differential stabilization, were also explored by factor analysis. The factor analysis and other results provide a basis for task evaluations. Task evaluations, comparison with previous studies, a consideration for future studies, and conclusions will be offered in the following sections.

Task Evaluations

Table 5 abstracts task characteristics revealed by earlier analyses. Examining this table, it can be noted that tasks are organized into four groups along lines suggested by the factor analysis. Number Facility (NF), Vertical Addition (Nv), Horizontal Addition (Nh), and Number Comparison (NC) constitute the first group which was identified with the first factor (F1). The second group is composed of Visualization (V) and Grammatical Reasoning (GR) which were identified with the second factor (F2). The third group is made up of Perceptual Speed (PS) and Speed of Closure (SC) measures which were identified with the third factor (F3). A fourth and last group is made up of the single Flexibility of Closure (FC) task. The FC task had low commonality with other tasks (.24), substantial stable reliability (0.88), and therefore substantial reliable "specificity". This specificity suggests defining FC as a separate factor with its loading equivalent to its reliability (0.88). For each group, the task loadings, stable periods for statistical measures, and stabilized reliabilities are also given. The factor groups provide collections of tasks which may be evaluated together using their abstracted characteristics. Evaluations, given below, will follow group organization.

> Factor 1 Group. Nv has both greatest reliability and factor loading of the members of this group. It evidences differential stability over Days 3 - 15 and is surpassed only by Nh, which had unstable variances. Both NF and NC appeared to obtain stability of means and variances slightly earlier in training, but both involve 180 second trials vice 120 seconds for Nv. Altogether Nv appears the choice from this group of tasks.

Table 5. Summary of Tasks Characteristics

Factor Group	Task	Factor Loading	Stabilized Days			Stabilize Reliability
			Means	Variances	Rel. Corr.	
	NF	0.85	9 - 13	3 - 13	9 - 13	0.83
	Nv	0.89	12 - 15	10 - 15	3 - 15	0.92
1	Nh	0.82	4 - 15	Unstable	1 - 15	0.79
	NC	0.71	7 - 15	10 - 15	3 - 15	0.77
	V	0.88	Unstable	Unstable	6 - 13	0.66
2	GR	- .68	2 - 15	3 - 15	6 - 15	0.87
	PS	0.88	9 - 13	1 - 15	7 - 13	0.84
3	SC	0.68	11 - 13	1 - 13	2 - 13	0.77
4	FC	0.88	3 - 13	1 - 13	3 - 13	0.88

Factor 2 Group. GR is the only member of this bipolar factor group which exhibits stability of means and variances. This is unfortunate as the opposition of spatial (V) and verbal (GR) tasks suggests a "cognitive style" factor as noted earlier. GR can be recommended as the only stable member of this group.

Factor 3 Group. PS has the largest loading reliability and earliest stabilization of means and variances. SC obtains differential stability earlier, but obtains overall stability later than PS. Hence, PS can be recommended as the representative of this group.

Factor 4 Group. FC is the only member of this group and is recommended by its overall stability.

Overall, the Vertical Addition (Nv), Perceptual Speed (PS), Grammatical Reasoning (GR), and Flexibility of Closure (FC) tasks may be recommended as the result of the evaluation.

Comparison with Previous Studies

The results from the current study may be compared with earlier investigations employing the same paradigm (Seales, et al., 1980; Carter, et al., 1981). Seales, et al. reported on a 10 minute test, a sequence of four arithmetic operation subtasks (addition, subtraction, multiplication, and division), which appeared to meet all stability criteria from the first day with a reliability of 0.941. In the present investigation, the NF, Nv, and Nh arithmetic tasks involved only the addition operation and respectively were 3, 2, and 2 minutes in duration with reliabilities of 0.83, 0.92, and 0.79. It might be expected that the reliabilities of the NF, Nv, and Nh tasks would have been of the order of 0.85 for a three minute task and 0.76 for a two minute task based on the Seales, et al. results and the Spearman-Brown Formula (Winer, 1971). Only the Nv results are out of line with these estimates ($p < .05$) with a greater than expected reliability of 0.92. The required periods for overall stabilization in the present investigation appear somewhat excessive, initially, but shortening the task length by factors of 3 to 5 may provide better assessment of the transitions to stability than the 10 minute block of the Seales, et al. (1980) task. The present investigation, in addition, employed more sophisticated statistical methodologies than Seales, et al., (1980). In any case, the present investigation indicates that the NF and Nv tasks require only 24 and 22 minutes of practice before they would be suitable for repeated measures applications.

The results of Carter, et al. (1981) are comparable with the current results for the Grammatical Reasoning (GR) Task. In their study, Carter, et al. reported that GR met all stability criteria after four daily (60 second) administrations with a stable reliability of $r = 0.82$. This investigation found that all stability criteria were met after six daily (90 second) administrations with a stable reliability of $\underline{r} = 0.87$. Perhaps due to the sharpened statistical methodologies, the period to meet all criteria was again somewhat lengthened although only 9 minutes total appears necessary even in the present study. The reliability of $\underline{r} = 0.87$ found in the present study is exactly what would be estimated from the Carter, et al. results and application of the Spearman-Brown Formula (Winer, 1971).

Altogether, the present results are comparable with those found in earlier studies with the same paradigm. More sensitive statistical methods are suspected of somewhat extending the required period for overall stability; however, required practice durations were not practically changed. The comparability of Spearman-Brown adjusted reliabilities across investigations supports the view that the tasks are differentially stabilized, as such stability is an assumption of the method. The stability of the adjusted stabilized reliabilities recommends, at least for arithmetic and grammatical reasoning tasks, estimation of reliabilities either by the Spearman-Brown Formula or by graphical methods (Bittner & Carter, 1981).

A Consideration for Future Studies.

The present investigation evaluated tasks in the units of measurement employed in the original research of Moran and Mefferd, (1959) and Carter and Sbisa (1981). In terms of numbers accomplished in a test period, the task scores are typical of a breadth of measures used to assess cognitive abilities and skills (cf, Ekstrom, et al., 1976). Transformations of scores were not examined despite repeated evidence suggesting their use. Of the nine measures examined, six (67%) were found with correlations between day number and log-transformed variances (viz, NF, V, Nv, Nh, NC, GR) and two (V and Nh) were apparently unstable over this study's duration. Prediction of increasing variances with trials, it is noteworthy, has been made by Jones (1972) for the class of tasks exemplified in this study. This prediction may be made from assumptions that (1) learning increases the rate of task processing and (2) individuals tend to retain their relative differences. The increase in variances over trials, usually paralleling the means, suggests scaling transformations with negative power strengths such as logarithmic, square root, etc. (Tukey, 1957). Other recently developed statistical methodology, involving conjoint measurement (Cliff, 1973, pp 475-476) and multidimensional-scaling (Carroll & Arabie, 1980, pp 629-630), might also prove of value for linearization. In any case, the selection of method of transformation or scaling is an empirical one which would require examination of the results in terms of the statistical requirements for repeated measures applications. Consideration of the use of transformation and scaling methods to improve the behavior of task scores appears desirable in future evaluations.

Conclusions

Two basic conclusions may be made based upon the results of this investigation: First, Vertical Addition (Nv), Perceptual Speed (PS), Grammatical Reasoning (GR) and Flexibility of Closure (FC) tasks may be recommended for repeated measures application subsequent to sufficient practice for stability. Second, the use of transformations and scaling techniques should be considered in future investigations of task stabilization.

References

Alluisi, E. A. Sustained performance. In (E. A. Bilodeau & I. McD. Bilodeau, (Eds.), Principles of skill acquisition. New York: Academic Press, 1969, 59-101.

Alvares, K. M., & Hulin, C. L. Two explanations of temporal changes in ability-skill relationships: A literature review and theoretical analysis. Human Factors, 1972, 14, 295-308.

Baddeley, A. D. A three minute reasoning test based on grammatical transformation. Psychonomic Science, 1968, 10, 341-342.

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

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

Bittner, A. C., Jr. Averaged correlations between parallel measures: Reliability estimation. (Research Report No. NBDL 81R013) Naval Biodynamics Laboratory, New Orleans, LA., 1981. (b)

Bittner, A.C., Jr., & Carter, R. C. Repeated measures of human performance: A bag of research tools. In J. C. Guignard (Ed.) Proceedings of the International Workshop on Research Methods in Human Motion and Vibration Studies, New Orleans, September, 1981, in press. (Also published as Naval Biodynamics Laboratory Research Report No. NBDL-81R011).

Campbell, D. T., & Stanley, J. C. Experimental and Quasi-Experimental designs for research. Chicago: Rand McNally, 1963.

Carroll, J. D., & Arabie, P. Multidimensional scaling. Annual Review of Psychology, 1980, 31, 607-649.

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, 13-17 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.

Carter, R. C., & Sbisa, H. Human performance tests for repeated measures: Alternate forms for eight tests by computer. New Orleans, LA. Unpublished manuscript, 1981. (Available from Naval Biodynamics Laboratory, New Orleans, LA. 70189.

Cliff, N. Scaling. Annual Review of Psychology, 1973, 24, 473-506.

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

Ekstrom, R. B., French, J. W., Harmon, H. H., & Dermen, D. Manual for kit of factor referenced cognitive tests. Princeton, N.J.: Educational Testing Service, 1976.

French, J. W. (Ed.) Manual for kit of selected tests for reference aptitude and achievement factors. Princeton, N. J.: Educational Testing Service, 1954.

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 Factors Society, Rochester, N. Y., October, 1981, 342-346.

Harmon, H. H. Modern factor analysis (3rd ed.). Chicago: University of Chicago Press, 1976.

Horn, J. L. State trait and change diversions of intelligence, British Journal of Educatinal Psychology , 42, 1972, 159-185.

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. (NBDL Monograph No. M-0001), New Orleans, LA: Naval Biodynamics Laboratory, 1980.

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

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

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, October, 1980, 344-348.

Moran, L. J., & Mefferd, R. B. Repetitive psychometric measures. Psychological Reports, 1959, 5, 269-275.

Moran, L. J., Kimble, J. P., Jr., & Mefferd, R. B. Repetitive psychometric measures equating alternate forms. Psychological Reports, 1964, 14, 335-338.

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

Seales, D. M., Kennedy, R. S., & Bittner, A. C., Jr. Development of Performance Evaluation Tests for Environmental Research (PETER): Arithmetic Computation. Perceptual and Motor Skills, 1980, 51, 1023-1031.

Steiger, J. H. Tests for comparing elements of a correlation matrix. Psychological 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.

Thomas, D. J., Majewski, P. L., Ewing, C. L., & Gilbert, N. S. Medical qualification procedures for hazardous-duty aeromedical research. AGARD Conference Proceedings No. 231. Nuilly-Sur-Seine, France: AGARD, 1978, A-3:1-13.

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

Tukey, J. W. The comparative anatomy of transformations. Annals of Mathematical Statistics, 1957, 602-632.

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


FILMED
7-8