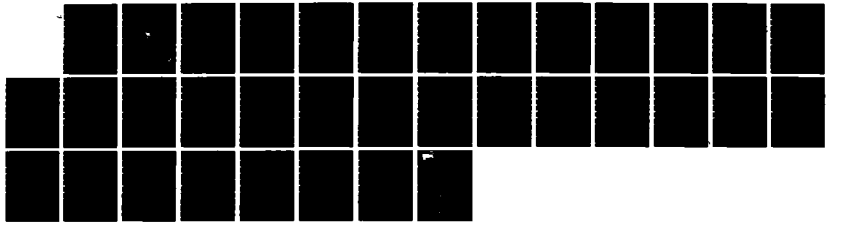
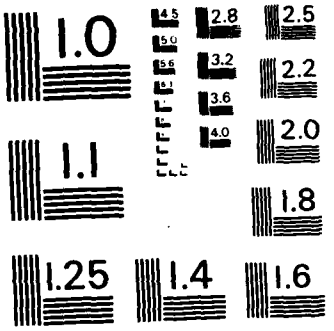
AD-A139 338 MASSED PRACTICE: DOES IT CHANGE THE STATISTICAL PROPERTIES OF PERFORMANCE TESTS?(U) NAVAL BIODYNAMICS LAB NEW ORLEANS LA M KRAUSE ET AL. JUN 83 NBDL-83R005 1/1

UNCLASSIFIED F/G 5/10 NL





NBDL-83R005

MASSED PRACTICE: DOES IT CHANGE THE STATISTICAL PROPERTIES OF PERFORMANCE TESTS?

Michele Krause and Jeffrey C. Woldstad

June 1983

NAVAL BIODYNAMICS LABORATORY
New Orleans, Louisiana

NBDL - 83R005

MASSED PRACTICE: DOES IT CHANGE THE STATISTICAL PROPERTIES OF PERFORMANCE TESTS?

Michele Krause

and

Jeffrey C. Woldstad

June 1983

Bureau of Medicine and Surgery
Work Unit Nos. MF 5852400E-0002 and M093004-0002

Approved by

Channing L. Ewing, M.D.
Chief Scientist

Released by

Captain L. E. Williams, MC USN
Commanding Officer

Naval Biodynamics Laboratory
New Orleans, Lousisiana

SUMMARY PAGE

PROBLEM

Repeated trials on a task are frequently required for assessing training procedures or experimental treatments. Limited time, money, or availability of research subjects often result in the need to give a substantial number of trials on a task within a short period of time. However, in many laboratories repeated measures are traditionally separated by 24 hours or more to reduce the chances of fatigue, interference, or other factors introducing undesirable error variance. Massing practice is an obvious alternative to distributing it, particularly when time constraints exist. However, massed practice is only a desirable alternative if the resulting test scores maintain the statistical properties required for repeated measures analysis.

FINDINGS

Paper-and-pencil and computerized versions of traditional human performance tests were examined under massed practice conditions. Many of the tests had been shown to have high reliabilities and to meet the statistical requirements for repeated measures applications under distributed practice conditions in earlier studies at our laboratory. The tests were: Grammatical Reasoning, Pattern Comparison, Purdue Pegboard, Aiming, Spoke, Maze Tracing, Code Substitution, Arithmetic, Stroop, and Memory Scanning. Although more time was required for task stabilization in most cases, all of the paper-and-pencil tasks retained high reliabilities under massed practice conditions, except Pattern Comparison and Maze Tracing. The latter appeared to have unequivalent alternate forms. Computer adaptations of task failed to maintain the statisiical properties required for repeated measures analysis.

RECOMMENDATIONS

It is recommended that distributed practice with trials separated by 24 hours or more be used whenever feasible. If massed practice is required tasks should be chosen which have been shown to have high reliability and which meet the statistical requirements for repeated measures experimentation. It is expected that once computer tasks are refined they too will lend themselves to massed practice administration when required.

The authors wish to thank Richard Irons and Timothy Whitten for their reliable computer programming/maintenance support. Special thanks to Robert Carter and Alvah Bittner, for sharing their data analysis expertise.

The volunteers used in this study were recruited, evaluated, and employed in accordance with the procedures specified in Secretary of the Navy Instruction Series 3900.39 and Bureau of Medicine and Surgery Instruction Series 3900.6. These instructions are based upon voluntary consent, and meet or exceed the provisions of prevailing national and international guidelines.

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

UNCLASSIFIED

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

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NBDL-83R005	2. GOVT ACCESSION NO. AD-A139338	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* Massed Practice: Does it Change the Statistical Properties of Performance Tests?		5. TYPE OF REPORT & PERIOD COVERED Research Report 6. PERFORMING ORG. REPORT NUMBER NBDL-83R005
7. AUTHOR(s) Michele Krause and Jeffrey C. Woldstad		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 M0933004-0002 MF 58 524 02E-0002
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Medical Research & Development Command Bethesda, MD 20014		12. REPORT DATE June 1983 13. NUMBER OF PAGES 32
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)*

Approval 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	Computerized Testing	Purdue Pegboard
Massed Practice	Memory Scanning	Aiming, Spoke,
Distributed Practice	Grammatical Reasoning	Maze Tracing, Coding,
Performance Testing	Pattern Comparison	Arithmetic, Stroop

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

Repeated trials on a task are frequently required for assessing training procedures or experimental treatments. Limited time, money, or availability of research subjects often result in the need to give a substantial number of trials on a task within a short period of time. However, in many laboratories repeated measures are traditionally separated by 24 hours or more to reduce the chances of fatigue, interference, or other factors which introduce undesirable error variance. Massing practice is an obvious alternative to

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

UNCLASSIFIED

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

distributing it, particularly when time constraints exist. However, massed practice is only a desirable alternative if the resulting test scores maintain the statistical properties required for repeated measures analysis. Paper-and-pencil and computerized versions of traditional human performance tests were examined under massed practice conditions. Many of the tests had been shown to have high reliabilities and to meet the statistical requirements for repeated measures applications under distributed practice conditions in earlier studies at our laboratory. The tests were: Grammatical Reasoning, Pattern Comparison, Purdue Pegboard, Aiming, Spoke, Maze Tracing, Code Substitution, Arithmetic, Stroop, and Memory Scanning. Although more time was required for task, stabilization in most cases, all of the paper-and-pencil tasks retained high reliabilities under massed practice conditions, except Pattern Comparison and Maze Tracing. The latter appeared to have unequivalent alternate forms. Computer adaptations of task failed to maintain the statistical properties required for repeated measures analysis. It is recommended that distributed practice with trials separated by 24 hours or more be used whenever feasible. If massed practice is required tasks should be chosen which have been shown to have high reliability and which meet the statistical requirements for repeated measures experimentation. It is expected that once computer tasks are refined they too will lend themselves to massed practice administration when required.

MASSED PRACTICE: DOES IT CHANGE THE STATISTICAL PROPERTIES OF PERFORMANCE TESTS?

INTRODUCTION

Parallelism of measurements is an assumption underlying repeated measures experiments (Jones, 1972; Lord & Novick, 1968; Winer, 1971). To ensure parallelism, repeated measures are usually separated by several hours at the least, and normally by 24 hours or more. Such long intervals between tests are necessary to avoid fatigue, proactive interference, and difficulty sustaining subjects' motivation. Given this, experiments which call for few measurements per subject are feasible. However, experiments requiring many repeated measures become impractical and often impossible. There are several reasons why an extended study is undesirable. First, the internal validity of experiments may be affected by extraneous events other than the experimental variables occurring between measurements (Campbell & Stanley, 1963, p.4). The probability of this occurrence increases as the number of measurements and the amount of time between measurements increases. Differential loss of respondents from the comparison groups, and maturation of subjects and apparatus are two examples of extraneous events enumerated by Campbell and Stanley which jeopardize internal validity of prolonged repeated measures experimentation. In addition, repeated measures experimentation is relatively expensive in terms of both experimenter and subject time. Nevertheless, the need for repeated measures experimentation exists. Therefore, it is worthwhile to investigate procedures which will yield comparable, reliable parallel measurements.

Lord and Novick (1968) and Jones (1980) have identified the statistical properties that tests should possess before they are used in repeated measures experimentation. Briefly, they define the requirements as: (1) constant or linearly increasing means across repeated measurements, (2) unchanging variances, and (3) differential stability. Differential stability, as outlined by Jones (1980), indicates that subjects' relative rank order is not changing and consequently intersession correlations remain constant. That is, the task should measure the same ability each time it is used. In addition, a task must have sufficient definition (Jones, 1980). Task definition is indicated by the averaged correlation across differentially stable trials. (See Bittner, 1981; Bittner, Dunlap, & Jones, 1982; Dunlap, Jones, & Bittner, 1983, for a justification of averaging correlations.)

Several human performance tests which meet these statistical criteria have been identified (Harbeson, Bittner, Kennedy, Carter, & Krause, 1983; Kennedy, Carter, & Bittner, 1980). These tests were examined using distributed trials of repeated measurement. The current study utilized a sample of those tests, for purposes of investigating whether massed practice yields results comparable to those obtained with distributed practice (i.e., results obtained when repeated measures were collected at daily intervals across several separate testing sessions). Both traditional apparatus and paper-and-pencil versions and newly programmed computer versions of classical tests were investigated. Comparable results between the two practice schedules would indicate that massed rather than distributed practice could be given to stabilize scores so that parallel repeated measures data could be collected.

The purpose of this study was to determine whether people perform comparably when given massed as opposed to distributed practice on tests. Tests which met the statistical requirements for parallel repeated measures when practice was distributed (i.e., trials were separated by > 24 hours) were given mass-practice to determine whether the desired statistical properties of the tests were again obtained.

METHOD

Experiment 1

Subjects

Seventeen Navy enlisted men between the ages of 18 and 25 were subjects for this experiment. All subjects were volunteers for environmental research experiments and met the health qualifications described by Thomas, Majewski, Ewing, and Gilbert (1978).

Apparatus and Task Descriptions

Six tests of cognitive, spatial, and motor ability were used in this study: Grammatical Reasoning, Pattern Comparison, Purdue Pegboard, Aiming, Spoke, and Maze Tracing. Each task is described below.

Grammatical Reasoning. This task, modeled after Baddeley's (1968), meets the statistical requirements for repeated measures testing (Carter, Kennedy, & Bittner, 1981). The Grammatical Reasoning test provides a measure of "higher mental processes" (Baddeley, 1968). Subjects were asked to decide whether a statement accurately described the relative position of two letters printed to the right of that statement. A typical item would look like:

A is preceded by B BA T F

The subjects were instructed to put a slash through the "T" if the statement was true and a slash through "F" if the statement falsely described the letter positions. Half of the statements were in the active voice (e.g., B follows A) and half passive (e.g., B is followed by A). Additionally, half were negative (e.g., A does not precede B) and half were positive statements (e.g., A precedes B). Twenty-four alternate forms, each with 32 items, were generated by a FORTRAN program (see Carter & Sbisa, 1982, for the program listing). The score recorded was the number correct minus the number incorrect for a 60 second administration.

Pattern Comparison. This test of perceptual speed was found to be suitable for repeated measures experimentation (Klein & Armitage, 1979; Shannon, Carter, & Boudreau, 1981). The object of this task was to determine whether two patterns were the same or different. A typical "different" trial looked like:

Subjects were instructed to write an "S" on the dashed line if the patterns were the same and a "D" if they were different. Subjects were given 144 total problems and 2 minutes to do as many as they could. The score was the number correct minus the number incorrect.

Purdue Pegboard. This is a test of finger dexterity designed by Science Research Associates, Inc. (Tiffin, 1968). Subjects were instructed to place cylindrical (2.5 mm in diameter) pegs into sequential holes until all were filled, or until the maximum time limit of two minutes was reached.

Aiming. This is a test of fine manipulative ability and is described more fully by Fleishman and Ellison (1962). The subject was required to make one dot in each of a series of very small circles (3 mm in diameter), working as quickly and as accurately as possible. The score was the number of dots correctly placed in 2 minutes.

Spoke. This task, which measures speed of lower arm movement, was fashioned after the Reitan Trail Making Test (Form A). Investigations indicate that this task is suitable for repeated measures use (Bittner, Lundy, Kennedy & Harbeson, 1982). The display sheets (43cm x 28cm) contained 32 circular targets arranged concentrically around a central circular target. Each target was 9.5mm in diameter and located 120.6mm from the central target. Distance from the center of one target to an adjacent target was 25.4mm. A number was displayed in the center of each target. The subject was required to alternately tap the stylus on the center target and each of the numbered circles (i.e., 0, 1, 0, 2, ... 0, 32). The score recorded was time to completion.

Maze Tracing. Ekstrom, French, Harman, and Dermen (1976) identify this task as loading on a spatial scanning factor. It measures the ability to find a path through 24 interconnected mazes. Variations of the original forms of this test were generated by Shannon (personal communication, 1982). The score was the number of blocks completed within 2 minutes.

Procedure

Testing was conducted on five consecutive weekdays, between the hours of 7:30 and 11:30 in the morning. Six tables, each with one test on it, were located around a large room. Subjects rotated from one table to the next in a different random order on each day until they had completed the full cycle.

After each cycle, the order in which subjects took each test was randomized. Eight replications of each test were administered on Day 1, followed by four replications on Days 2 - 5. Subjects were tested in two groups of five and one group of seven.

Experiment 2

Subjects

The subjects were 14 Navy enlisted men between the ages of 18 and 25. All subjects were volunteers for the same environmental research program specified in Experiment 1 and met the health qualifications.

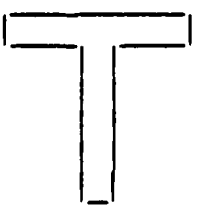
Apparatus

Testing equipment included APPLE II PLUS® microcomputers connected to and controlled by a NESTAR CLUSTER/ONE MODEL A® central networking system. This system provided for simultaneous testing on six APPLE® computers and automatic data storage from each testing station. Each computer was equipped with 64K memory, an interval timing clock (Mountain Hardware Inc.), and an APPLE® language card. Stimulus was presented on 13-inch screens. Four Zenith® monitors and two Quasar® color TV's were used. In addition to the APPLE® keyboard, a numeric keypad (Advanced Business Technology, Inc.), standard APPLE® paddles, and a three-button box built in-house served as response devices. Hence, input to the subject was visual while manual responses were required.

Task Descriptions

Six well-known tests of mental functioning, Code Substitution, Math, Stroop, Memory Scanning, Grammatical Reasoning, and Pattern Comparison, were used in this study. All tests were programmed in Applesoft Basic® language, for implementation on the APPLE® microcomputers. Program listings for these tests are available from the authors upon request. Computer adaptations of the six tests used in this study are described in detail below.

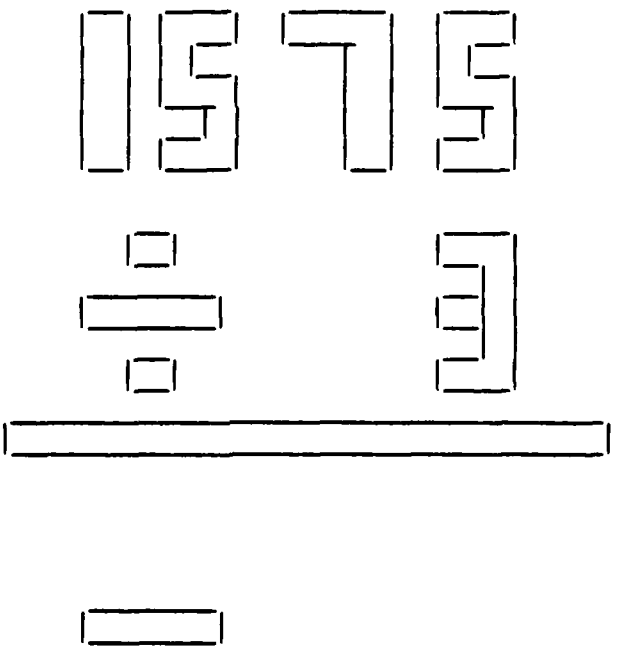
Code Substitution. This test is conceptually the same as that on the Wechsler Adult Intelligence Scale-Revised (1980). It has been found to meet the statistical criteria necessary for repeated measures testing (Pepper, Kennedy, Bittner, & Wiker, 1981). Pairs of letters and numbers were presented to the subjects and their task was to respond with the appropriate digit when a code letter was presented alone. Nine letters, generated randomly by the computer, were paired with the digits one (1) through nine (9). During a trial, one of the nine letters would print in the top three-quarters of the screen, above the digit-code pairs, and would remain until the subject pushed one of the keys on a nine-key numerical pad. Throughout the three minute task, the coded pairs remained the same. The score was the number of correct responses. A sample of the stimulus display, as it appeared on the screen is:



CODE	A	M	N	T	S	R	Q	V	X
DIGIT	5	1	4	2	8	7	9	3	6

Arithmetic. This test of arithmetic computation is similar to the Number Facility tests described by Ekstrom, French, Harman, and Dermen (1976). Its statistical reliability and stability indicate that it is an appropriate test to use in repeated measures testing (Seales, 1980). This task included

addition, subtraction, multiplication, and division problems. Within blocks of four, each type of problem was randomly presented once. In an attempt to keep difficulty levels equivalent, addition was restricted to 3-by-3 and 3-by-2 problems, subtraction to 3-by-3 only, division to 1-by-3, and 1-by-4, and multiplication to 1-by-2. The task was to perform the computation mentally and enter the answer on a 13-key numerical pad. A marker on the screen indicated where the subject should start keying in responses. Once a response was typed it could be changed by pushing an "erase" key or entered, allowing for the next problem to appear on the screen. Numbers were graphed across the center of the screen using the low resolution graphics mode, and each measured approximately 2.5cm x 2.5cm. A sample division problem is:



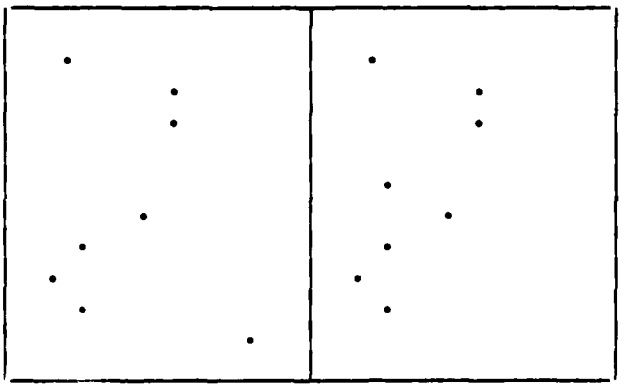
The other three types of problems looked similar, except that responses were entered from right to left on addition, subtraction, and multiplication problems rather than from left to right as in the division problems. Problems were presented consecutively, for two minutes, with approximately two seconds between the subject's entering a response and the presentation of another. The score was the number of correct responses for problems that were started within the two minute time frame.

Stroop. This test represents one of several versions of a serial verbal task involving interference, which was designed by Stroop (1935,1938). The version used in this experiment was similar to one found to have the statistical characteristics necessary for repeated measures testing (Harbeson, Krause, Kennedy, & Bittner, 1982). Subjects were instructed to respond to either a word or a color in this task. There were three conditions: black-white (BW), color-word (CW), and color-color (CC). In the BW condition, the words "RED", "BLUE", and "GREEN" were presented on the screen in random order, and in black-and-white. Subjects were instructed to push buttons which corresponded

to the word appearing on the screen. The words, "RED", "BLUE", and "GREEN" were also used in the CW and CC conditions, but were written in one of the three colors so that the color might or might not have matched what the word said. In the CW condition, subjects were instructed to ignore the color and respond only to the word. Just the opposite was requested in the CC condition; subjects were asked to ignore the word and respond to the color the word was written in. Letters were graphed across the center of the screen, using the high resolution mode, and were approximately 3.0cm x 4.0cm in size. Subjects responses were input on a three-button box, with one being red, one blue, and one green. In addition, the letters 'R', 'B', and 'G' were typed above the corresponding button. Approximately one second lapsed between trials within a condition, and each condition lasted 45 seconds. Each condition was presented once within a sitting, but in random order each time. The order remained the same across subjects. The score for this test was number correct minus a portion (.33) of the incorrect responses.

Memory Scanning. Sternberg's (1966, 1975) information processing task was used as a model in programming this test. A "target" of from one to four digits was presented, immediately followed by a single "probe" digit. The subjects task was to respond positively or negatively, depending on whether the "probe" digit was one of the "target" digits. Stimuli were randomly generated numbers from zero through nine. The numbers were graphed across the center of the screen, using the low resolution graphics mode, and were 2.5 cm x 2.5 cm each. Six trials at each target size (one through four) were randomly presented, for a total of 24 trials per sitting. Average reaction time to each trial was recorded. The usefulness of this test for repeated measures applications is discussed by Carter, Kennedy, Bittner, and Krause (1981), and Carter and Krause (1983).

Pattern Comparison. The object of this task was to determine whether two patterns presented on the screen (one in the left half and one in the right half) were the same or different. High resolution graphics were used to present patterns, each composed of approximately eight dots, on each half of the screen. Points were randomly selected using a random number generator, and dots were plotted in rapid succession, with the left pattern appearing slightly before the pattern on the right side of the screen. An example of a "different" trial is:



Subjects responded by pushing one of two keys on the keyboard marked "S" and "D". Approximately one second after each response, another set of patterns would appear on the screen. The task lasted for three minutes. The score was the number correct minus the number of incorrect responses.

Grammatical Reasoning. This test was programmed according to Baddeley's (1968) specifications, as described in Experiment 1. Thirty two sentences were randomized and presented, sequentially, to subjects as quickly as they could respond to whether the statement on the screen was true or false. The intertrial interval was about one second. Responses were made on the buttons, one marked "T" and the other "F", of standard APPLE® paddles. The task ended automatically at the end of two minutes. The score was the number correct minus the number of incorrect responses.

Procedure

A few days prior to the first day of the experiment, subjects were shown the laboratory set-up and the basics of operating the apparatus. Additionally, the instructions were reviewed and subjects were given practice trials on each test. The experiment was conducted over a five-day period. Within a session, subjects moved from one station to another completing each of the tests which were housed in six separate booths. On Day 1, testing was completed within 150 minutes, and five replications were run on each test. Three replications on each test, requiring about 75 minutes, were given on the four subsequent days. Ten subjects were tested between the morning hours of 8:00 and 11:00; the remaining four subjects were tested between 12:30 and 3:30 in the afternoon.

RESULTS

General Analysis Method

The initial stage of analysis for both experiments included checking the data for outlying and missing points. Some subjects' data were eliminated from the analysis on this basis. The total number of subjects included in the analysis for each test is indicated in parentheses, following each test name. An additional step in the initial stage of analysis was to calculate for each test the correlations between group means and variances to determine whether a transformation of the raw data was necessary. Transformations used are specified.

Secondly, Days X Trials repeated measures analysis-of-variances (ANOVAs) were computed for the means and, separately, for the jackknife variance estimates (Carter & Bittner, 1982) for each test. This provided for examining the days effect, and trials-within-a day effect, as well as their interaction. Intersession correlations were analyzed sequentially by Steiger's (1980) method, using the approach described by Bittner and Carter (1981), to determine whether at any point in practice they ceased to change significantly.

A summary of the test administration times, scores recorded for each test and stability results are outlined in Table 1 for Experiment 1 and Table 2 for Experiment 2.

Experiment 1

Grammatical Reasoning (N = 17)

The resulting means, standard deviations, and correlations for this test are listed in Table 3. As indicated, group mean scores showed a linear trend over days, after the initial four trials were dropped ($F(4,64) = 8.50$, $p < .001$). Means across the four trials within each day were relatively homogeneous ($F(12,192) = .89$, $p > .50$). Variances (listed in italics along the diagonal in Table 3) were unchanged across days ($F(5,80) = .69$, $p > .60$) and trials ($F(3,48) = .50$, $p > .65$). The Days by Trials interaction was also nonsignificant ($F(15,240) = .75$, $p > .70$). Steiger (1980) analysis method indicated that intertrial correlations were stable across the last nine trials ($\chi^2(35) = 35.38$, $p > .45$), with an averaged reliability of .83. The delayed stabilization appeared to be due to two unusually high correlations (.95) between trials 23, 24, and previous trials. Otherwise, as indicated by Table 3, intertrial correlations were relatively homogeneous across trials 10-24.

Pattern Comparison (N = 17)

A correlation of .71 between the means and standard deviations suggested a log transformation of the raw data. The transformed group means increased linearly over days subsequent to the first day ($F(3,48) = 6.19$, $p = .001$, overall; $F(1,16) = 8.08$, $p = .012$, linear). The linear component accounted for 88% of the total variation. Means remained relatively constant across the last three trials of each day ($F(2,32) = 2.15$, $p > .10$). The Days X Trials interaction was significant ($F(6,96) = 3.98$, $p < .01$). The interaction is due to changes in subjects' relative intertrial performances with increased practice. In the initial experimental days, practice effects were apparent across trials within a day. However, by later days, performance on the second trial was essentially the same as performance on the third and fourth trials. As indicated in Table 4, intertrial correlations were essentially homogeneous across the final seven trials ($\chi^2(20) = 24.05$, $p > .24$). The averaged reliability across stable trials was .81. Overall, correlations tended to be higher for adjacent trials and declined as trials became more separated in time.

Purdue Pegboard (N = 17)

Group means were homogeneous across days, after dropping the initial four trials on Day 1 ($F(4,64) = .54$, $p > .70$). Additionally, trials within a day remained constant ($F(3,48) = .67$, $p > .50$). The Days X Trials interaction was significant ($F(12,192) = 2.27$, $p < .01$), however, it failed to remain statistically significant after Day 1 was dropped from consideration ($F(9.144) = 1.75$, $p > .08$). Variances remained constant across all days ($F(5,80) = .78$, $p > .57$) and across trials within each day ($F(3,48) = 1.12$, $p > .35$). The Days X Trials interaction for variances was also statistically insignificant ($F(15,240) = 1.60$, $p > .07$). Intertrial correlations were stable across only the last three trials ($\chi^2(2) = 2.62$, $p > .26$), with an averaged reliability of .84. Table 5 indicates that both same-day and cross-day intertrial correlations were intermittently high and low, with no obvious pattern. This suggests that the relative order of subjects' performances continued to change quite drastically, until the last three trials.

Aiming (N = 17)

Means remained constant across days ($F(4,64) = .60$ $p > .65$) with the first four trials excluded. Means within a day, across trials, showed statistically significant change that was mainly due to a significant linear trend ($F(3,48) = 21.28$, $p < .001$, overall; $F(1,16) = 41.62$, $p < .001$, linear). When the initial trial each day was dropped, the significant linear trend persisted, accounting for 95% of the significant change ($F(2,32) = 5.36$, $p < .01$, overall; $F(1,16) = 10.46$, $p < .005$, linear). The linear trend in means within each testing day was essentially the same across days, as indicated by a nonsignificant Days X Trials interaction ($F(8,128) = .47$, $p < .80$). Jackknife variance estimates remained relatively constant across days and trials within each day (respectively, $F(5,80) = 1.12$, $p > .35$ and $F(3,48) = 1.81$, $p > .15$). In addition, there was a nonsignificant Days X Trials interaction ($F(15,240) = .61$, $p > .85$). Intertrial correlations across Trials 16-24 were stable ($\chi^2(35) = 42.49$, $p > .18$) with an averaged reliability of .82. The intertrial correlations fluctuated randomly, with more low correlations occurring as trials were more separated by time (Table 6).

Spoke (N = 17)

Group means across the five experimental days were significantly different ($F(5,80) = 19.23$, $p < .001$). Means within each day, across trials, also fluctuated significantly ($F(3,48) = 18.37$, $p > .001$). Group means did remain constant, however, across the last four days, with the first trial of each day excluded ($F(3,40) = .65$, $p > .58$). Within each day, means were constant across the last three trials ($F(2,32) = .02$), $p > .97$). There was a significant Days X Trials interaction, however, ($F(6,96) = 3.52$, $p < .01$), a definite nonlinear trend on Day 1 with an increasingly linear trend across trials toward later days. Variances remained constant across days ($F(5,80) = .73$, $p > .60$), and trials ($F(3,48) = 2.19$, $p > .10$). In addition, their interaction was nonsignificant ($F(15,240) + 1.50$, $p > .10$). Intertrial correlations were stable across the last eight trials ($\chi^2(35) = 44.73$, $p > .10$) and the averaged reliability was .86. Table 7 indicates that correlations failed to stabilize across trials 10-24 because of occassional low correlations. However, intertrial correlations of trials earlier than 10 with later trials were more consistently low, particularly as they were more separated in time.

Maze Tracing (N=17)

An ANOVA on daily group means indicated a highly significant Days effect when means were blocked across trials ($F(5,80) = 72.64$, $p < .001$). There was no interpretable trend in the means; first through fifth order effects were highly significant. This indicated that there was an erratic pattern in the means across days. The Trials effect, blocked across days, was also highly significant ($F(3,48) = 15.23$, $p < .001$). In addition, the Days X Trials interaction was also significant ($F(15,240) = 35.95$, $p < .001$). Figure 1 shows the unusual pattern in the means that underlies the highly significant results. As can be seen in the graph, means increase across the first 15 trials, then drop drastically on Trial 16, and again increase linearly throughout the remainder of the experiment. Fifteen alternate forms of this test were used on the initial 15 trials of this experiment. On Trials 16-24 the first nine alternate forms were reiterated. The resulting means indicated

that the alternate forms were probably not equivalent; a linear increase of difficulty across forms is suspected. This test is dropped from futher discussion since any findings would be overshadowed by the nonequivalence of the alternate forms.

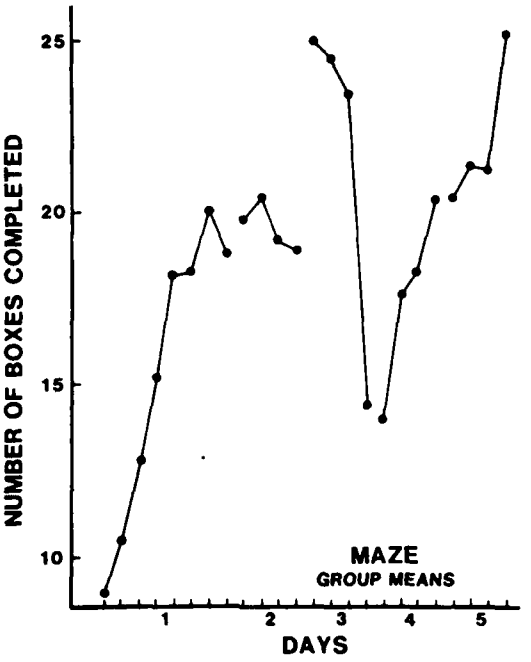


Figure 1. Maze Tracing: Mean number completed for 24 replications over 5 days (n = 17).

Experiment 2

Code Substitution (N = 13)

The ANOVA computed on the means revealed a significant Days effect ($F(4,48) = 13.70$, $p < .001$). A substantial amount of this effect, 97%, was attributable to the linear component ($F(1,12) = 43.35$, $p < .001$). The overall Trials effect was nonsignificant ($F(2,24) = 1.35$, $p > .27$), as was the Days X Trials interaction ($F(8,96) = 1.41$, $p > .20$). Thus, while within-day trial means remain relatively constant, the group means across the five days indicate a steady increase with practice. An ANOVA calculated on the jackknife variance estimates indicated nonsignificant Days and Trials effects, as well as a nonsignificant interaction ($F(4,48) = 1.25$, $p > .30$, $F(2,24) = 2.23$, $p > .13$, $F(8,96) = .85$, $p > .56$, respectively). Jackknife variances, therefore, remained constant across both Days and Trials. Intersession correlations were differentially stable across trials 11-15 ($\chi^2(9) = 13.31$, $p > .14$), however, the average correlation across the stable trials (task definition) was extremely low ($r = .26$). Hence, although the means increased linearly, and variance estimates remained constant over trials, this task lacks reliability. As can be seen in Table 9, the intertrial correlations for this task were generally low, even for trials given in succession. There is no obvious reason for this finding, although the computer program might be suspected.

Arithmetic (N = 12)

A significant Days effect appeared in the ANOVA on group means ($F(4,44) = 5.85$, $p < .001$). This effect was 99% linear ($F(1,11) = 14.13$, $p < .003$), with none of the other components reaching statistical significance. The Trials main effect, on the other hand, was insignificant ($F(2,22) = 1.49$, $p > .24$), indicating that within-day trials remained constant. The Days X Trails interaction was also insignificant ($F(8,88) = 1.59$, $p > .13$). A comparable ANOVA on the jackknife variance estimates indicated a significant Days main effect when all five days were considered ($F(4,44) = 2.68$, $p < .05$). Both the Trials main effect and the Days X Trials interaction were insignificant ($F(2,22) = .06$, $p > .94$, and $F(8,88) = .34$, $p > .94$, respectively). Dropping the initial day (Trials 1-3) from the analysis brought about insignificant Days, Trials, and interaction effects ($F(3,33) = 1.05$, $p > .38$, $F(2,22) = .06$, $p > .94$, and $F(6,66) = .49$, $p > .80$, respectively). Variances, therefore, remained constant across Days 2-5 of all experiment. Intersession correlations were differentially stable across all 15 trials, as evidenced by the Steiger (1980) analysis ($\chi^2(104) = 96.32$, $p > .69$). The averaged reliability of .55 indicated poor task definition when all 15 trials were included; however, there was a trend toward increased reliability as the initial trials were dropped from consideration. Averaged reliability across Trials 11-15 was .77. This relatively low intertrial reliability, when compared to the paper-and-pencil version, may be due to the fact that this test includes all four numerical operations, which vary in difficulty for most individuals. If the amount of time spent on each type of numerical operation is variable this could contribute to the overall instability of the test. Recall that each type of problem was presented at random, once within each block of four problems. Given that the test ended after approximately two minutes, it was possible for the most difficult type of problem for any one subject (i.e., the type of problem that required the most time for arrival at an accurate solution) to outnumber the easier problem on one day but possibly not the next day. This variability could be eliminated by changing the Math test to include only one type of problem (e.g., addition) rather than all four types represented here.

Stroop (N = 9)

An ANOVA computed on the means for all days revealed a significant Days X Trials interaction ($F(8,64) = 2.13$ $p < .05$). This significant interaction impeded interpretation of the m[illegible]ffects although neither the Days nor Trials effect was significant ([illegible] = 1.75, $p > .16$ and $F(2,16) = 1.60$, $p > .23$, respectively). When the [illegible] was dropped from the analysis, the Days X Trials interaction was [illegible]gnificant ($F(6,48) = 1.58$, $p > .17$). Again, the Days and Trials main [illegible]were also nonsignificant ($F(3,24) = .80$, $p > .51$ and $F(2,15) = .20$, $p > .$ [illegible], respectively). Hence, after the first testing day, means remained unchanged. Jackknife variance estimates also remained constant over all Trials and Days as indicated by an ANOVA which wncluded all observations. The statistical values for Days and Trials main effects and the Days X Trials interaction were repectively $F(4,32) = .09$, $p > .98$, $F(2,16) = .29$, $p > .74$, and $F(8,64) = 1.98$, $p > .06$. This indicates that variances for the Stroop CW score remain constant across practice trials. Steiger analysis on the intertrial correlations revealed an averaged reliability of .28 across the 15 trials, although the correlations were differentially stable ($\chi^2(104) = 97.09$, $p > .67$). When initial trials were

dropped and only Trials 12-15 were examined, the averaged reliability rose to .49. There was an inherent problem in the computer version of the Stroop test, which surfaced while subjects were doing the test and which may account for the unreliable intertrial correlations. After the instructions appeared on the screen and the test began, the subjects frequently forgot whether they were supposed to attend to the colors the words were written in, or attend to the words, disregarding the color. That is, they were often unable to remember the instructions pertaining to the specific task once the instructions left the screen and the test began. Although only the color-word condition was scored, the subjects were asked to do both types of tasks, at different times. Confusion could be eliminated by administering only one type of Stroop task.

Memory Scanning (N = 13)

The ANOVA on group means for all observations showed a significant main effect for Days ($F(4,48) = 5.32$, $p < .01$). Thirty-six percent of the total sums-of-squares was attributed to the linear component ($F(1,12) = 6.82$, $p > .02$), while 61% was quadratic ($F(1,12) = 17.09$, $p > .01$). The Trials main effect was not statistically significant ($F(2,24) = .69$, $p > .50$), and neither was the Days X Trials interaction ($F(8,96) = .59$, $p > .78$). Excluding the initial testing day resulted in stable, unchanging means for the remaining Days ($F(3,36) = .60$, $p > .61$) and Trials ($F(2,24) = 1.11$, $p > .34$). The Days X Trials interaction was again nonsignificant ($F(6,72) = .58$, $p > .74$). Analysis of the jackknife variance estimates showed that they remain constant across all Trials ($F(2,24) = .14$, $p > .87$) and all Days ($F(4,48) = 1.18$, $p > .32$) of the experiment. There was no significant interaction between Days and Trials ($F(8,96) = .71$, $p > .68$). Intertrial correlations were differentially stable across Trials 6 - 15 ($\chi^2(44) = 54.70$, $p > .12$) with an averaged reliability of .78. Table 10 shows that between Trials 1-15 and all other trials there was an occasional high or low intertrial correlation which kept the correlations from reaching statistical equivalence. Other than that, the intertrial correlations are relatively homogeneous throughout.

Pattern Comparison (N = 10)

Group means remained constant across all observations, as indicated by the analysis-of-variance. The statistical values for the Days, Trials, and interaction effects were repectively $F(4,36) = 1.91$, $p > .12$, $F(2,18) = .33$, $p > .72$, and $F(8,72) = 1.20$, $p > .30$. An ANOVA on the jackknife variance estimates revealed that they also remained essentially unchanged across the course of the experiment (Days: $F(4,36) = 1.90$, $p > .13$; Trials: $F(2,18) = 1.53$, $p > .24$; Days X Trials: $F(8,72) = .48$, $p > .86$). Intertrial correlations were differentially stable across Trials 2 - 15 ($\chi^2(90) = 105.85$, $p > .12$), with an averaged reliability of .47. Reliabilities fell off, reaching approximately .27 when only the last four trials were considered. Table 11 indicated that intertrial correlations were moderate to low, with no apparent pattern, throughout the matrix.

Grammatical Reasoning (N = 13)

All data analysis on this test include Days 2-5; Day 1 was excluded because the data was lost in computer transmission. An ANOVA on the group means showed that the main effects were nonsignificant, with $F(3,36) = .51$, p

> .67 for Days and $F(2,24) = .26$, $p > .77$ for the Trials effect. In addition, the Days X Trials interaction was also nonsignificant ($F(6,72) = .56$, $p > .75$). Therefore, grouped means remained constant across the final four experimental days. An ANOVA on the jackknife variance estimates suggested that there was no statistically significant change in the variances across Days ($F(3,36) = 1.6$, $p > .91$) or Trials ($F(2,24) = .36$, $p > .69$). Additionally, their interaction was nonsignificant ($F(6,72) = .69$, $p > .66$). Steiger analysis of the intertrial correlations indicated that Trials 8 - 15 were differentially stable ($\chi^2(27) = 31.24$, $p > .26$), with an averaged reliability of .85. As indicated in Table 11, intertrial correlations were moderate to high, and relatively homogeneous throughout. However, two low correlations (between Trials 12 and 7, and 14 and 7) apparently prevented the matrix of intertrial correlations from being statistically homogeneous prior to Trial 8.

DISCUSSION

The results, summarized in Tables 1 and 2, indicate that when mass practiced, most tests either lose or take longer to achieve the statistical properties required of tests used for repeated measures applications. Interesting comparisons can be made between the massed and distributed-practice results, and likewise between paper-and-pencil and computer massed-practice results. These comparisons will be discussed in turn.

Daily group means and variances for mass practiced computer tasks generally stabilize early in relation to distributed practice results for the equivalent paper-and-pencil tests (Table 2). However, the correlational results for mass practiced computer tasks are disappointing. Overall, the intertrial correlations and consequent averaged reliabilities indicate a lack of task definition. That is, when these particular tests are subjected to massed practice, the attribute(s) being measured change from one trial to the next. One exception appears to be Grammatical Reasoning, which reaches an acceptable level of reliability (.85), although it takes longer to attain stability when mass practiced. The computer version of Grammatical Reasoning appears to yield a higher averaged reliability (for the same amount of test time) when mass practiced than our paper-and-pencil version does when practice is distributed (.85 vs. .80).

The apparatus and paper-and-pencil massed practice and distributed practice results are compared in Table 1. Generally, the group means take longer to stabilize when mass practiced. Variances, on the other hand, remained constant across trials, except for Pattern Comparison, whose variances never stabilized. Intertrial correlations took longer to stabilize in all cases when mass practiced, however, the majority of the tasks reached an acceptable averaged reliability when they finally became homogeneous (all above .81).

In summary, paper-and-pencil tasks examined, except Pattern Comparison, retain statistical properties required for repeated measures applications when mass practiced. Both group means and intertrial correlations require substantially longer to stabilize, however, when mass practiced than when practice is distributed. With the exception of Grammatical Reasoning, the computer tasks failed to reach an acceptable level of reliability, and therefore are not suited for use in their present forms. Grammatical Reasoning in its computer form is not as reliable as the paper-and-pencil version, but might be acceptable for some applications because of convience.

A change that may improve the reliability of the computer tasks is to reduce (ideally remove) the opportunity for subjects to make ambiguous or unintentional responses.

Paper-and-pencil tests have a long history and have continually been refined and improved over time. Computer adaptations of traditional tests are relatively new, and therefore it is reasonable that time and effort may need to be expended before they have the reliability and construct validity of their traditional counterparts. If a serious effort is made to continually scrutinize and improve computer tasks, in the future we may have a way of testing abilities that is superior to the traditional testing approach. Computers may potentially act to reduce experimental error by functioning as consistant, reliable test administrators. An added advantage is that a computer can score and analyze tests quickly and accurately. These factors are merely a few which contribute to the promise that computerized testing holds for the future. In order for computerized testing to provide meaningful results, however, we must improve the reliability of computer testing procedures, hardware, and software.

Evidence reviewed leads us to conclude that distributed practice should be used when possible. In situations where economic or other constraints dictate that mass practice is necessary, well established paper-and-pencil and apparatus tests which yield high reliability should be used. A sufficient number of trials (more than 20 in most cases) should be given to ensure that stability is reached prior to repeated measures use. New computer tasks should be scrutinized carefully for factors leading to unreliability and instability prior to their use.

REFERENCES

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

Bittner, A. C., Jr. Averaged correlations between parallel measures: reliability estimation. Proceedings of the 8th Psychology in the DoD Symposium (USAFA-TR-82-10). Colorado Springs: USAF Academy, April 1982, 321-327.

Bittner, A. C., Jr., & Carter, R. C. Repeated measures of human performance: a bag of research tools (Research Report No. NBDL-81R011). Naval Biodynamics Laboratory, New Orleans, 1981. (NTIS No. AD A113954)

Bittner, A. C., Jr., Dunlap, W. P., & Jones, M. B. Averaged correlations with differentially-stable variables: Fewer subjects required for repeated measures. Proceedings of the 26th Annual Meeting of the Human Factors Society. Santa Monica, CA: Human Factors Society, 1982, 349-353.

Bittner, A. C., Jr., Lundy, N. C., Kennedy, R. S., & Harbeson, M. M. Performance Evaluation Tests for Environmental Research (PETER): Spoke tasks. Perceputual and Motor Skills, 1982, 54, 1319-1331.

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

Carter, R. C., & Bittner, A. C., Jr. Jackknife for variance analysis of multifactor experiments (Research Report No. NBDL-82R013). Naval Biodynamics Laboratory, New Orleans, May 1982. (NTIS No. AD A121760)

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., Kennedy, R. S., Bittner, A. C., Jr., & Krause, M. Item recognition as a Performance Evaluation Test for Environmental Research. Proceedings of the 24th Annual Meeting of the Human Factors Society. Santa Monica, CA: Human Factors Society, 1980, 340-344. Also, (Research Report No. NBDL-81R008) Naval Biodynamics Laboratory, New Orleans, July 1981, 47-50. (NTIS No. AD A111296)

Carter, R. C., & Krause, M. Reliability of slope scores for individuals. (Research Report No. NBDL-83R003) Naval Biodynamics Laboratory, New Orleans, April 1983. (NTIS No. AD A130252)

Carter, R. C., & Sbisa, H. E. Human performance tests for repeated measurements: alternate forms of eight tests by computer (Research Report No. NBDL-82R003). Naval Biodynamics Laboratory, New Orleans, January 1982. (NTIS No. AD A115021)

Dunlap, W. P., Jones, M. B., & Bittner, A. C., Jr. Average correlations vs. correlated averages. Bulletin of the Psychonomic Society, 1983, 21, 213-216.

Ekstrom, R. B., French, J. W., Harman, H. H., & Derman, D. Manual for kit of factor-referenced cognitive tests. Princeton, New Jersey: Educational Testing Service, 1976.

Fleishman, E. A., & Ellison, W. E., Jr. A factor analysis of five manupulative tests. Journal of Applied Psychology, 1962, 46, 96-105.

Harbeson, M. M., Bittner, A. C., Jr., Kennedy, R. S., Carter, R. C., & Krause, M. Performance Evaluation Tests for Environmental Research (PETER): Bibliography. Perceptual and Motor Skills, 1983, 57, 283-293.

Harbeson, M. M., Krause, M., Kennedy, R. S., & Bittner, A. C., Jr. The Stroop as a performance evaluation test for environmental research. The Journal of Psychology, 1982, 111, 223-233.

Jones, M. B. Stabilization and task definition in a performance test battery. (Final Report on Contract No. N0023-79-M-5089, Monograph No. NBDL-M001). Naval Biodynamics Laboratory, New Orleans, October 1980. (NTIS No. AD A099987)

Jones, M. B. Individual differences. In R. N. Singer (Ed.), The Psychomotor Domain. Philadelphia: Lea and Fabinger, 1972.

Kennedy, R. S., Carter, R. C., & Bittner, A. C., Jr. A catalogue of Performance Evaluation Tests for Environmental Research. Proceedings of the 24th Annual Meeting of the Human Factors Society. Santa Monica, CA: Human Factors Society, 1980, 334-348. Also, (Research Report No. NBDL-80R008) Naval Biodynamics Laboratory, New Orleans, July 1981, 8-12. NTIS No. AD A111296)

Klein, R. & Armitage, R. Rhythms in human performance: $1^1/^2$-hour oscillations in cognitive style. Science, 1979, 204, 1326-1328.

Krause, M. Paper-and-pencil and computerized performance tests: Does the medium make a difference? New Orleans: Naval Biodynamics Laboratory, in preparation.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Pepper, R. L., Kennedy, R. S., Bittner, A. C., Jr., & Wiker, S. F. Performance Evaluation Tests for Environmental Research (PETER): code substitution test. Proceedings of the 7th Psychology in the DoD Symposium (USAFA-TR-80-12). Colorado Springs: USAF Academy, April 1980, 451-457. Also, (Research Report No. NBDL-80R008) Naval Biodynamics Laboratory, New Orleans, July 1981, 13-19. (NTIS No. AD A111296)

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

Shannon, R. H. Personal communication, 1982.

Shannon, R. H., Carter, R. C., & Boudreau, Y. A. A systematic approach to battery development and testing within unusual environments. In J. C. Guignard & M. M. Harbeson (Eds.), Proceedings of the International Workshop on Research Methods in Human Motion and Vibration Studies. New Orleans: Naval Biodynamics Laboratory, September, 1981, in preparation.

Steiger, J. H. Tests for comparing elements of a correlation matrix. Psychological Bulletin, 1980, 87, 245-251.

Sternberg, S. High speed scanning in human memory. Science, 1966, 153, 652-654.

Sternberg, S. Memory scanning: new findings and current controversies. Quarterly Journal of Experimental Psychology, 1975, 27, 1-32.

Stroop, J. R. Studies of interference in serial verbal reactions. Journal of Experimental Psychology, 1935, 18, 643-662.

Stroop, J. R. Factors affecting speed in serial verbal reactions. Psychological Monograph, 1938, 50, 38-48.

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. Neuilly-Sur-Seine, France: AGARD, 1978, A-3: 1-13.

Tiffin, J. Manual for Purdue Pegboard. Chicago, Illinois: Science Research Associates, 1968.

Wechsler, D. Manual for the Wechsler Adult Intelligence Scale - Revised. New York: The Psychological Corporation, 1980.

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

Table 1: Experiment 1 - Paper and Pencil/Apparatus Tasks

TEST	MEASURE	TEST TIME	MASSED PRACTICE STABILITY RESULTS				DISTRIBUTED PRACTICE STABILITY RESULTS**			
			MEANS	VARIANCES	INTERTRIAL CORR.	AVG. REL.	MEANS	VARIANCES	INTERTRIAL CORR.	AVG. REL.
MAZE	# OF BLKS COMPLETED	2 MIN	SIG DAYS & TRIALS	(UNEQUIVALENT ALTERNATE FORMS)						
SPOKE-C	LN [TIME TO COMP]	2 MIN	9 - 24Δ	1 - 24	16 - 24	*.86	2 - 15	2 - 15	1 - 15	.80
AIMING	# OF TAPS	2 MIN	5 - 24	1 - 24	16 - 24	.82				
PERDUE PEGBD	TIME TO COMPLETE	2 MIN	9 - 24	1 - 24	22 - 24	.84				
GRAM. REAS.	C - E	1 MIN	5 - 24	1 - 24	16 - 24	.83	5 - 15	1 - 15	5 - 15	.80
PATTERN COMP.	L_n[C-E]	2 MIN	13 - 24Δ	DO NOT STABILIZE ACROSS DAYS OR TRIALS	18 - 24	.81	10 - 15	10 - 15	10 - 15	.89

* Stability results for Massed Practice are reported by Trial Number (Eight successive trials given Day 1, four trials given Days 2-5)

** Stability results for Distributed Practice are reported by Session Number (one trial given per session)

Δ With first trial each day dropped

Table 2: Experiment 2 - Computer Tasks

TEST	MEASURE	TEST TIME	MASSED PRACTICE STABILITY RESULTS *				DISTRIBUTED PRACTICE STABILITY RESULTS**			
			MEANS	VARIANCES	INTERTRIAL CORR.	AVG. REL.	MEANS	VARIANCES	INTERTRIAL CORR.	AVG. REL.
CODE SUB	# CORR	2 MIN	1 - 15	1 - 15 L	11 - 15	.26	8 - 15	8 - 15	8 - 15	.75
ARITH.	# CORR	3 MIN	1 - 15	4 - 15 L	1 - 15	.55 → .77	9 - 15	1 - 15	1 - 15	.94
STROOP-CW	# CORR - .33(I)	45 SEC	4 - 15	1 - 15	1 - 15	.28 → .49	6 - 15	6 - 15	6 - 15	.85
MEMORY SCAN	TIME/C ITEM	24 ITEMS	4 - 15	1 - 15	6 - 15	.78				
PATTERN COMP.	# CORR - # INC.	3 MIN	1 - 15	1 - 15	2 - 15	.47	10 - 15	10 - 15	10 - 15	.89
GRAM. REAS.	# CORR - # INC.	1 MIN	1 - 15	1 - 15	8 - 15	.85	5 - 15	1 -15	5 - 15	.80

* Stability results for Massed Practice are reported by Trial Number (Three successive trials reported per day)
** Stability results for Distributed Practice are reported by Session Number (one trial given per session)
L = Majority of significant days effect was linear

Table 3: Cross-session correlations, daily group means and standard deviations for Experiment 1 - GRAMMATICAL REASONING*

		24											13
6.29	1	.43	.26	.40	.51	.28	.58	.38	.62	.49	.63	.63	.65
10.12		.63	.45	.46	.64	.51	.77	.51	.63	.52	.56	.69	.80
9.70		.59	.46	.42	.67	.51	.78	.54	.71	.58	.67	.72	.83
10.88		.52	.47	.55	.69	.52	.76	.59	.70	.62	.69	.74	.82
11.29		.68	.71	.75	.89	.77	.84	.77	.87	.86	.87	.92	.94
10.94		.76	.77	.76	.88	.77	.87	.77	.78	.76	.80	.87	.86
12.82		.71	.73	.75	.83	.76	.88	.71	.71	.71	.79	.85	.84
12.65		.69	.70	.75	.88	.80	.88	.73	.80	.82	.82	.89	.90
13.65		.78	.79	.84	.88	.91	.87	.89	.93	.88	.88	.92	.93
15.00		.74	.72	.71	.92	.76	.87	.76	.79	.80	.80	.90	.97
13.76		.78	.83	.80	.91	.83	.87	.90	.85	.83	.86	.92	.95
13.70		.79	.79	.73	.88	.75	.85	.82	.81	.79	.84	.90	.95
15.18		.78	.74	.79	.93	.81	.88	.83	.88	.86	.88	.95	9-11
14.23		.81	.81	.88	.95	.83	.86	.86	.88	.89	.90	8-58	7-09 1
16.06		.70	.77	.87	.84	.77	.81	.82	.90	.95	8-43	6-65	.78
15.18		.70	.76	.87	.86	.83	.78	.78	.88	7-82	7-88	.85	.81
15.53		.70	.67	.80	.80	.81	.77	.87	8-09	8-84	.82	.91	.79
16.18		.80	.86	.87	.84	.87	.78	7-25	8-29	.75	.79	.69	.60
15.65		.89	.85	.81	.86	.85	8-49	8-50	.85	.74	.79	.67	.64
16.12		.84	.85	.88	.85	7-48	8-54	.92	.87	.76	.76	.70	.64
16.29		.81	.83	.88	8-74	9-75	.92	.92	.93	.82	.79	.73	.61
16.35		.83	.90	9-71	7-74	.92	.86	.89	.91	.74	.73	.69	.55
16.59		.90	9-10	7-22	.88	.89	.82	.87	.89	.82	.84	.81	.61
15.94	24	8-07	8-47	.93	.93	.87	.84	.87	.92	.74	.76	.69	.49
		8-09	.94	.94	.87	.80	.81	.85	.86	.72	.82	.72	.61 12
		12											1

*Group means are along the left margin, standard deviations along the diagonal (in italics) and correlations within the upper and lower triangles

Table 4: Cross-session correlations, daily group means and standard deviations for Experiment 1 - PATTERN COMPARISON*

	24											13
1.72	1	.34	.54	.35	.47	.27	.44	.35	.36	.47	.49	.46	.38
1.79		.27	.38	.32	.48	.42	.52	.53	.47	.53	.58	.51	.54
1.81		.39	.46	.37	.54	.59	.52	.54	.52	.60	.58	.52	.58
1.84		.38	.61	.43	.63	.57	.63	.59	.58	.68	.70	.75	.63
1.86		.57	.72	.58	.66	.59	.73	.57	.57	.70	.67	.64	.60
1.88		.67	.83	.73	.77	.69	.79	.68	.72	.81	.78	.73	.68
1.86		.56	.74	.64	.73	.72	.81	.72	.73	.87	.84	.80	.76
1.85		.49	.66	.49	.58	.47	.59	.37	.45	.66	.54	.64	.48
1.89		.54	.65	.63	.65	.70	.70	.68	.74	.74	.81	.74	.79
1.86		.47	.68	.59	.71	.82	.74	.77	.77	.88	.89	.86	.89
1.90		.59	.74	.65	.75	.81	.79	.74	.77	.87	.87	.89	.79
1.92		.64	.74	.70	.74	.76	.88	.74	.77	.90	.85	.80	.80
1.91		.66	.70	.77	.83	.90	.84	.73	.80	.84	.86	.84	-09 1
1.91		.59	.72	.70	.77	.81	.80	.78	.86	.82	.80	-09	-11
1.94		.58	.71	.71	.74	.82	.78	.87	.90	.90	-10	-07	.82
1.94		.66	.81	.77	.80	.90	.85	.84	.86	-09	-06	.75	.55
1.92		.64	.71	.82	.77	.83	.84	.93	-08	-08	.78	.78	.78
1.94		.54	.67	.71	.75	.77	.81	-09	-08	.81	.67	.74	.82
1.93		.81	.82	.89	.88	.86	-07	-08	.89	.81	.58	.71	.83
1.94		.80	.78	.86	.89	-07	-08	.92	.92	.88	.69	.74	.76
1.95		.86	.88	.91	-11	-12	.84	.80	.82	.72	.41	.53	.79
1.97		.92	.85	-11	-07	.56	.79	.82	.79	.66	.66	.77	.67
1.95		.84	-12	-10	.83	.59	.83	.70	.71	.73	.66	.62	.47
1.94	24	-13	-09	.88	.84	.75	.93	.88	.85	.86	.69	.70	.66
		-08	.87	.82	.75	.78	.94	.86	.82	.72	.52	.64	.63 12
		12											1

*Group means are along the left margin, standard deviations along the diagonal (in italics) and correlations within the upper and lower triangles

Table 5: Cross-session correlations, daily group means & standard deviations for Experiment 1 - PEGBOARD*

51.35	1	-.38	-.40	-.46	-.19	-.19	-.25	-.28	-.10	-.19	-.33	-.28	-.14
47.82		-.19	-.27	-.20	.11	.04	-.06	.02	.09	-.05	.08	.19	.29
47.18		-.17	-.20	-.04	.18	.21	.20	.24	.34	.10	.22	.34	.16
43.65		-.05	-.04	.22	.17	-.02	.40	.24	.08	.32	.05	.47	-.04
44.35		-.00	.06	.09	-.19	-.13	.10	.18	-.22	.02	-.15	.23	-.14
43.06		.10	.01	.36	.14	.16	.37	.24	.14	.52	.19	.42	.05
41.18		.06	.01	.35	.12	.10	.45	.22	.10	.34	-.04	.37	-.12
43.29		-.09	-.07	.09	.03	.01	.16	.21	-.01	.40	-.03	.24	-.10
42.41		.25	.53	.39	.21	.32	.55	.57	.46	.39	.40	.79	.22
43.59		.12	.16	.40	.37	.12	.39	.60	.26	.54	.47	.60	.26
45.59		.22	.16	.37	.56	.47	.15	.54	.56	.46	.78	.54	.66
41.94		.23	.40	.53	.50	.37	.68	.70	.64	.65	.62	.83	.35
43.23		.29	.28	.37	.67	.39	.21	.34	.73	.32	.85	.63	3~93
42.23		.36	.53	.57	.61	.50	.72	.69	.80	.51	.75	5~18	6~01
42.41		.28	.33	.49	.67	.39	.38	.63	.79	.58	2~98	6~44	.60
42.65		.06	.1.1	.33	.43	.19	.40	.53	.44	4~34	4~14	.74	.39
42.53		.26	.39	.48	.68	.52	.59	.52	3~12	4~09	.33	.23	-.29
42.06		.39	.56	.52	.49	.51	.65	3~80	4~65	.63	.06	.15	-.08
42.53		.61	.66	.74	.65	.65	6~10	3~98	.53	.67	.25	.14	-.19
44.29		.83	.75	.71	.67	4~56	4~72	.81	.60	.83	.17	.09	-.21
42.29		.57	.46	.68	3~29	4~01	.68	.64	.24	.61	.07	.16	-.15
41.00		.87	.77	5~50	6~95	.41	.39	.30	.25	.50	.11	-.04	-.46
42.52		.89	5~86	4~76	.55	.32	.56	.46	.56	.75	.18	.04	-.50
41.94	24	7~13	6~98	.53	.29	.11	.00	.19	-.08	.15	.45	.29	-.21
		6~51	.51	.67	.81	.55	.50	.50	.11	.60	.24	-.04	-.50

*Group means are along the left margin, standard deviations along the diagonal (in italics) and correlations within the upper and lower triangles

Table 6: Cross-session correlations, daily group means and standard deviations for Experiment 1 - AIMING*

		24											13
258.06	1	.67	.72	.76	.71	.71	.59	.66	.59	.78	.78	.81	.78
284.41		.40	.34	.25	.24	.35	.40	.28	.29	.36	.32	.56	.66
301.29		.50	.36	.42	.37	.41	.47	.42	.37	.58	.65	.64	.74
305.82		.67	.63	.63	.48	.66	.60	.60	.60	.68	.69	.74	.80
326.41		.65	.52	.47	.45	.55	.67	.61	.58	.70	.66	.67	.82
327.35		.73	.66	.57	.48	.57	.71	.68	.63	.76	.68	.68	.92
328.88		.56	.48	.42	.46	.49	.68	.58	.54	.65	.61	.60	.84
330.29		.44	.45	.49	.53	.51	.47	.46	.38	.62	.62	.57	.69
305.76		.64	.49	.47	.45	.52	.58	.56	.49	.55	.52	.58	.83
324.23		.74	.63	.60	.59	.60	.68	.64	.63	.74	.72	.80	.87
338.82		.77	.68	.62	.58	.70	.71	.73	.71	.71	.66	.75	.88
337.18		.78	.75	.73	.61	.77	.74	.73	.76	.75	.77	.88	.79
330.00		.82	.79	.72	.61	.73	.84	.83	.79	.84	.77	.82	46-18
334.23		.78	.78	.81	.76	.73	.75	.75	.82	.84	.84	35-45	53-53 1
333.23		.68	.66	.76	.77	.69	.73	.78	.73	.91	42-51	49-66	.58
338.12		.82	.81	.79	.80	.76	.85	.87	.86	44-84	39-40	.82	.66
306.70		.83	.89	.8C	.72	.83	.90	.89	42-21	43-64	.90	.76	.74
322.94		.87	.88	.79	.72	.87	.91	46-47	48-64	.92	.93	.79	.63
330.18		.78	.79	.68	.68	.76	39-41	48-29	.92	.87	.84	.72	.73
332.88		.83	.89	.86	.75	59-71	42-56	.91	.90	.81	.84	.79	.65
310.82		.76	.73	.85	53-64	53-23	.69	.65	.62	.58	.69	.66	.79
326.59		.86	.89	57-70	40-68	.60	.87	.83	.80	.74	.78	.75	.53
333.76		.92	59-57	42-11	.85	.71	.88	.91	.95	.90	.91	.83	.72
339.29	24	54-37	40-97	.92	.86	.65	.90	.88	.90	.86	.80	.76	.68
		43-34	.80	.86	.64	.49	.66	.75	.82	.90	.76	.65	.70 12
		12											1

*Group means are along the left margin, standard deviations along the diagonal (in italics) and correlations within the upper and lower triangles

Table 7: Cross-session correlations, daily group means and standard deviations for Experiment 1 - SPOKE*

		24											13
1.59	1	.43	.35	.30	.42	.51	.51	.54	.44	.43	.65	.57	.69
1.54		.35	.26	.23	.39	.39	.49	.54	.38	.44	.71	.67	.70
1.50		.48	.36	.27	.37	.42	.53	.57	.46	.46	.68	.69	.73
1.47		.48	.41	.33	.31	.40	.52	.58	.48	.53	.70	.68	.72
1.46		.43	.39	.35	.31	.41	.50	.58	.47	.55	.77	.67	.75
1.44		.47	.38	.39	.41	.50	.62	.70	.52	.56	.80	.79	.79
1.44		.63	.56	.48	.57	.63	.68	.61	.64	.66	.73	.67	.86
1.42		.55	.45	.49	.48	.52	.68	.72	.55	.55	.77	.77	.81
1.47		.49	.42	.41	.47	.51	.55	.58	.52	.58	.72	.65	.74
1.44		.68	.55	.53	.59	.64	.76	.72	.66	.62	.81	.75	.91
1.41		.68	.58	.61	.60	.68	.78	.88	.63	.67	.89	.94	.84
1.40		.76	.65	.63	.64	.72	.81	.83	.65	.67	.86	.86	.88
1.42		.73	.67	.70	.80	.80	.87	.84	.81	.79	.89	.85	-06
1.40		.74	.66	.72	.71	.77	.83	.93	.70	.78	.89	-08	-09 1
1.40		.67	.68	.77	.73	.74	.75	.84	.74	.85	-07	-07	.87
1.40		.78	.86	.87	.84	.85	.74	.79	.88	-09	-08	.90	.72
1.42		.83	.88	.85	.91	.92	.83	.83	-08	-07	.91	.77	.56
1.41		.85	.81	.83	.81	.89	.91	-08	-08	.95	.85	.79	.66
1.41		.89	.78	.80	.88	.92	-09	-08	.93	.88	.85	.82	.70
1.43		.93	.92	.90	.94	-09	-08	.82	.85	.83	.82	.72	.68
1.41		.84	.85	.90	-08	-09	.82	.92	.89	.86	.77	.65	.52
1.40		.85	.93	-09	-06	.79	.87	.85	.81	.73	.73	.70	.64
1.43		.93	-12	-07	.76	.87	.91	.85	.88	.87	.86	.78	.71
1.41	24	-12	-10	.84	.67	.83	.70	.88	.78	.78	.82	.80	.69
		-08	.94	.94	.74	.86	.85	.86	.83	.82	.86	.79	.74 12
		12											1

*Group means are along the left margin, standard deviations along the diagonal (in italics) and correlations within the upper and lower triangles

Table 8: Cross-session correlations, daily group means and standard deviations for Experiment 1 - MAZE TRACING*

		24											13
9.18	1	.13	.18	.11	.29	.50	.55	.28	.55	.52	.44	.55	.36	
10.76		.47	.35	.42	.54	.63	.66	.53	.68	.56	.65	.62	.63	
13.47		.43	.08	.22	.32	.59	.53	.48	.59	.37	.46	.39	.57	
15.23		.45	.31	.41	.45	.35	.57	.54	.62	.50	.45	.47	.68	
18.12		.65	.57	.64	.64	.40	.85	.71	.81	.64	.67	.65	.75	
18.23		.56	.48	.61	.57	.32	.70	.65	.73	.54	.56	.54	.52	
20.06		.67	.39	.64	.57	.44	.60	.65	.70	.44	.64	.52	.31	
18.82		.70	.58	.61	.73	.83	.74	.75	.79	.70	.78	.70	.42	
19.88		.81	.56	.69	.71	.71	.80	.81	.76	.66	.75	.72	.69	
20.53		.86	.63	.73	.75	.70	.83	.91	.77	.68	.81	.79	.71	
19.23		.81	.62	.71	.75	.64	.84	.87	.83	.67	.85	.80	.76	
19.06		.79	.63	.72	.70	.66	.85	.78	.87	.73	.92	.81	.80	
24.82		.68	.58	.67	.66	.46	.80	.67	.72	.64	.73	.77	5-65	13
24.41		.75	.67	.83	.86	.62	.80	.74	.79	.75	.86	6-76	1-78	1
23.47		.83	.65	.83	.79	.65	.81	.80	.79	.67	3-92	2-05	.61	
14.53		.63	.85	.66	.80	.76	.84	.71	.83	2-53	2-27	.75	.41	
13.70		.77	.70	.73	.79	.75	.89	.83	2-99	2-56	.70	.74	.33	
17.76		.94	.74	.83	.81	.73	.82	2-51	3-69	.80	.54	.70	.42	
18.41		.76	.78	.72	.81	.67	3-18	3-21	.91	.68	.38	.63	.33	
20.47		.67	.59	.51	.65	3-48	2-54	.76	.72	.54	.48	.54	.27	
20.53		.80	.86	.89	3-97	3-50	.75	.59	.61	.44	.59	.73	.52	
21.47		.91	.79	3-73	4-12	.82	.75	.63	.77	.71	.75	.75	.37	
21.23		.74	5-18	3-66	.92	.81	.65	.59	.74	.62	.62	.63	.40	
25.12	24	5-87	2-93	.92	.81	.75	.62	.58	.74	.63	.59	.65	.51	
		3-78	.91	.83	.78	.75	.62	.61	.78	.62	.57	.71	.55	12
		12											1	

*Group means are along the left margin, standard deviations along the diagonal (in italics) and correlations within the upper and lower triangles

Table 9: Cross-session correlations, daily group means and standard deviations for Experiment 2 - CODE SUBSTITUTION (upper diagonal and right margin) & ARITHMETIC (lower diagonal and left margin)*

	15														1	
		8-59	.50	-.04	.52	-.12	-.04	-.38	.30	.18	.46	.12	.26	.02	-.01	-.32	91.46	15
1	7.25	2-42	10-58	.32	.34	.19	.54	.11	.57	.13	.24	.35	.37	.46	.17	.10	93.00	
	7.83	.64	2-44	14-21	.04	.70	.73	.79	.80	.42	.51	.67	.67	.42	.46	.59	97.23	
	7.83	.76	.77	2-21	6-96	.20	.30	.16	.40	-.09	.14	-.04	-.04	.11	.16	.21	89.61	
	8.25	.78	.91	.75	2-93	11-10	.72	.79	.76	.49	.56	.60	.59	.32	.42	.59	92.92	
	9.42	.81	.78	.68	.88	2-50	9-07	.67	.83	.36	.27	.56	.50	.65	.75	.76	89.08	
	6.92	.68	.69	.79	.77	.83	3-15	11-37	.61	.25	.32	.47	.39	.27	.35	.61	88.23	
	8.08	.53	.68	.63	.53	.58	.45	3-06	10-77	.45	.66	.69	.73	.40	.53	.56	91.61	
	9.75	.77	.70	.80	.84	.85	.94	.43	3-57	8-92	.55	.74	.53	.36	.57	.43	87.15	
	8.92	.64	.79	.75	.66	.67	.62	.68	.51	3-78	7-69	.74	.69	.14	.12	.02	83.61	
	9.33	.36	.49	.62	.49	.55	.73	.47	.59	.68	3-98	7-05	.75	.64	.48	.51	85.54	
	8.67	.38	.38	.33	.55	.27	.17	.11	.26	.39	.34	3-80	8-74	.28	.43	.31	82.46	
	9.83	.58	.75	.79	.67	.60	.58	.45	.63	.70	.50	.23	2-85	6-59	.56	.59	84.92	
	10.00	.48	.58	.56	.46	.53	.72	.45	.52	.78	.64	.01	.44	3-69	7-25	.73	82.92	
	10.42	.82	.67	.63	.87	.91	.76	.45	.87	.56	.56	.47	.59	.40	3-58	11-76	79.69	1
15	9.83	.45	.62	.60	.58	.50	.54	.03	.53	.50	.16	.09	.72	.43	.36	3-69		
		1														15		

*Standard deviations are along the diagonal (in italics), means are along the right and left margins, and correlations are within the upper and lower triangles

Table 10: Cross-session correlations, daily group means and standard deviations for Experiment 2 - STROOP upper diagonal and right margin) - MEMORY SCANNING (lower diagonal and left margin)*

		15														1		
		2-21	.67	.80	.00	-.35	.14	.21	.21	-.23	.42	.32	.45	-.08	.29	.43	37.41	15
1	1.17	-19	3-19	.77	.27	-.28	.10	.23	.61	.05	.39	.38	.08	.13	.20	.19	36.97	
	1.16	.83	-19	3-60	.45	-.05	.23	.55	.58	.05	.59	.59	.54	.49	.34	.25	37.04	
	1.16	.87	.86	-20	3-67	.69	-.36	.86	.50	.50	.46	.30	-.02	.71	.30	-.11	37.74	
	1.07	.77	.48	.71	-21	3-38	-.18	.47	.05	.23	.46	-.02	.07	.57	.53	-.11	35.41	
	1.10	.88	.75	.90	.81	-20	2-09	-.37	-.03	-.09	.12	.49	.77	.25	.22	.24	36.71	
	1.10	.81	.61	.74	.66	.80	-19	3-15	.50	.58	.32	.33	.09	.69	.04	.20	36.41	
	1.03	.78	.62	.78	.67	.87	.83	-21	2-07	.39	.56	.58	.08	.56	-.19	.23	38.30	
	1.09	.85	.69	.80	.76	.88	.74	.91	-21	2-55	-.12	.52	-.01	.52	-.40	.40	37.33	
	1.09	.73	.77	.95	.62	.87	.66	.81	.81	-17	2-57	.44	.47	.44	.55	.17	36.45	
	1.06	.89	.69	.79	.80	.90	.90	.82	.86	.73	-15	3-34	.67	.57	-.06	.39	36.33	
	1.08	.84	.85	.94	.67	.92	.71	.84	.90	.95	.82	-15	2-90	.45	.36	.40	35.63	
	1.07	.81	.58	.81	.81	.88	.83	.72	.72	.72	.89	.77	-18	1-82	.18	.10	36.81	
	1.09	.80	.64	.80	.83	.88	.74	.75	.86	.78	.88	.85	.87	-17	2-99	-.18	36.63	
	1.09	.85	.73	.93	.79	.95	.82	.90	.90	.92	.89	.94	.89	.91	-22	6-35	32.78	1
15	1.11	.81	.52	.75	.76	.75	.76	.82	.77	.68	.74	.71	.72	.63	.80	-25		
		1														15		

*Standard deviations are along the diagonal, (in italics) means are along the right and left margins, and correlations are within the upper and lower triangles

Table 11: Cross-session correlations, daily group means and standard deviations for Experiment 3 - PATTERN COMPARISON (upper diagonal and right margin) & GRAMMATICAL REASONING (lower diagonal and left margin)*

	15														1	
1		5-26	.18	.68	-.05	.24	.65	.53	.44	.70	.45	.88	.67	.60	.61	.47	31.10	15
			3-82	.00	.38	-.26	.43	-.09	.50	.22	.32	.10	.01	.30	.47	.22	31.20	
				6-08	.35	.55	.46	.60	.60	.50	.52	.87	.56	.45	.47	.55	32.90	
					5-13	.11	.42	.17	.51	.21	.50	.18	.21	.04	.22	.45	31.50	
	3.70				2-65	3-99	.19	.19	.45	.19	.05	.41	-.18	-.07	.17	.12	33.20	
	4.47				.81	3-28	4-77	.69	.55	.85	.86	.69	.54	.75	.90	.82	33.50	
	4.39				.89	.92	2-47	5-78	.30	.69	.85	.64	.72	.81	.72	.88	33.10	
	4.38				.91	.84	.87	2-56	3-91	.65	.41	.48	.29	.26	.56	.60	32.00	
	3.63				.94	.82	.90	.86	2-37	4-98	.70	.66	.65	.68	.83	.86	32.80	
	3.39				.83	.74	.83	.71	.92	2-99	6-28	.60	.60	.85	.85	.90	30.80	
	4.78				.85	.88	.93	.82	.92	.91	3-14	4-98	.64	.67	.68	.56	33.10	
	4.01				.81	.78	.88	.72	.83	.87	.88	2-27	5-09	.59	.45	.70	28.80	
	4.93				.83	.71	.86	.66	.91	.92	.91	.91	2-65	6-34	.90	.76	29.60	
	4.24				.93	.92	.95	.92	.92	.80	.92	.88	.84	2-86	7-67	.81	30.00	
	3.40				.78	.76	.88	.69	.79	.72	.84	.90	.89	.89	2-47	9-96	29.90	1
15	4.23				.88	.81	.90	.81	.93	.87	.86	.87	.90	.90	.81	3-27		
		1														15		

*Standard deviations are along the diagonal, (in italics) means are along the right and left margins, and correlations are within the upper and lower triangles