CR-172038

PRELIMINARY EVALUATION OF A MICRO-BASED REPEATED MEASURES TESTING SYSTEM

Robert S. Kennedy, M.A., Ph.D.
Robert L. Wilkes, B.A., M.A.
Norman E. Lane, M.A., Ph.D.

Essex Corporation
1040 Woodcock Road, Suite 227
Orlando, Florida
(305) 894-5090

Jerry L. Homick, M.A., Ph.D.
National Aeronautics Space Administration
Houston, Texas

ESSEX ORLANDO
TECHNICAL REPORT

EOTR 85-1

ABSTRACT

Introduction. A need exists for an automated performance test system to study the effects of various treatments which are of interest to the aerospace medical community. The ethics and pragmatics of such assessment demand that repeated measures in small groups of subjects become the customary research paradigm. In such cases test stability, reliability-efficiency and factor structure take on extreme significance; in a program of study by the U.S. Navy, 80% of 150 tests studied failed to meet minimum metric requirements.

Methods. The "best" of these tests are being programmed on a portable microprocessor and administered along with tests in their original formats in order to examine their metric properties in the computerized mode. Twenty subjects have been tested over four replications on a 6.0 minute computerized battery (six tests) and which compared with five paper and pencil marker tests.

Results. All tests achieved stability within the four test sessions, reliability-efficiencies were high ($r>.707$ for three minutes testing) the computerized tests were largely comparable to the paper and pencil version from which they were derived. Two well-defined factors emerged from the 6.0 minute test battery.

Conclusions. This portable, inexpensive, rugged, computerized test battery can be employed and is recommended for study of the effects of drugs and environmental stress.

INTRODUCTION

Preface

Exotic work environments often include factors (ie., weightlessness, motion, fatigue, etc.) that disrupt performance. Furthermore, these settings are typically populated by limited numbers of highly critical workers. Some (e.g., Kennedy & Bittner, 1977) have observed that two connected concerns associated with the measurement of performance under such conditions are the lack of sensitive tests and a general unwillingness to expend the time and effort necessary to standardize such a test battery. It is tautological that the quality of data-based decision making is limited by the quality of the data on which the decision is made, and decisions directly reflect the adequacy of the assessment instruments and procedures employed in generating information. Helmstadter (1964) has emphasized the importance of carefully developed and administered tests in "providing the best information possible as a guide to decision making" (p. 32). Situations involving unique informational needs or atypical data collection settings dictate even greater attention to detail. Certain military and aerospace work environments obviously qualify for special attention.

Attention solely to the adequacy of the test battery may not satisfy all the demands of performance testing in exotic environments. Situational demands may dictate efficient and convenient procedures for data collection and storage. Time factors may be critical, necessitating rapid analysis of data and immediate feedback of results. These concerns suggest that innovative methods for data collection description and analysis must be explored.

In recent years there has been widespread interest in computerized performance tests. The Army, Navy, Air Force, Veterans Administration, Environmental Protection Agency, other agencies, and several universities have active programs. These research programs constitute valuable resources for the development of a computerized testing system. Thorne, Genser, Sing, and Hegge (1983) administered this Performance Assessment Battery (PAB) in a 72-hour sleep deprivation experiment. Eight subjects participated in a laboratory environment under high task load conditions. Performance, mood-activation and physiological measures were taken. The PAB was shown to be sensitive to changes in performance, with all tasks showing similar decrement patterns across time. A neurophysiological microprocessor test battery is being developed at the Air Force Aerospace Medical Research Laboratory (AFAMRL) to assess the effects of workload on operator performance. Tests are being implemented in software to be used by nontechnical personnel in field environments (O'Donnell, 1981). In addition, a subjective workload scale is also being developed (Reid, Shingledecker, Nygren & Eggemeier, 1981). The Learning Abilities Measurement Program (LAMP) at the Air Force Human Resources Laboratory (AFHRL) is investigating individual differences in cognitive abilities and information processing (Christal, 1981; Payne, 1982). Tests have been programmed on microcomputers in a laboratory with 30 automated testing stations. Approximately 30 tests have been developed under contract. Data have been collected on 24 tests and preliminary results are generally comparable to those in the literature (Christal, Payne, Weissmuller, & Anderson, 1982).

In the Appletox program, which is sponsored by the Environmental Protection Agency at the University of North Carolina at Chapel Hill, Eckerman and his colleagues are developing an automated test battery to detect the effects of

toxic substances on human performance. The primary test device is an APPLE II microcomputer. Tests identified by the cognitive experimental paradigm of J. B. Carroll (1980) have been selected for evaluation. Seven tasks have been implemented thus far. More tasks are in process, some data have been collected, and refinement of tasks and technical equipment is ongoing (Eckerman, Personal Communication, June 1981). According to Cramer (1982), NASA has also "identified a need for improved methods of assessing the effects of exotic environments on human performance." In particular, questions regarding the effects of space motion sickness and anti-motion drug treatment have been raised. Assessment of such environmental effects and potential remedies can best be accomplished with testing systems compatible with exotic environments. Automated and portable microprocessors capable of administering and storing performance measures and responses provide the obvious vehicle. Performance testing in hazardous situations with a scarcity of qualified subjects suggests the use of a repeated measures approach. Such an approach has been specifically recommended for research involving reduced sample size (Carter, Kennedy, & Bittner, 1981) and minimized exposure time (Bittner, Carter, Kennedy, Harbeson, & Krause, 1984). Furthermore, it has been emphasized (Kennedy & Bittner, 1977; Bittner, Carter, Kennedy, Harbeson, & Krause, 1984) that the individual tasks forming a test battery for use in a repeated measures paradigm require extensive evaluation and testing prior to application.

Background

A program designed to develop Performance Evaluation Tests for Environmental Research (PETER) was undertaken by the Naval Aerospace Medical Research Laboratory Detachment, New Orleans, Louisiana (Kennedy & Bittner, 1977, 1978). The purpose of this program was to develop a repeated measures test battery, effective in measuring human performance decrements over time, or in unusual work environments. The PETER paradigm was based on an "engineering approach" whereby specific criteria for test properties were established for tests in the battery to meet. Typically, prior to inclusion in a final battery or list of tests, a candidate task would be administered to a group of subjects through a series of 15 sessions performed on 15 successive working days. Data were collected across repeated measures and specific statistical criteria were applied to test and evaluate the potential candidate tasks. Particular emphasis was directed toward the assessment of test stability. Guilford (1965), strongly recommends in such efforts that "...As a general policy it would be desirable to establish the principles regarding what kinds of tests yield stable scores, with what kinds of populations and over what periods of time, and what kinds of tests do not" (p. 452).

More than 150 tests were studied for the final test battery (Bittner, Carter, Kennedy, Harbeson, & Krause, 1984). Tests qualifying as potential candidates were first determined to be appropriate for repeated measures assessment (ie., possess comparable alternate forms) and second, to measure mental work. Furthermore, selection of tests for study was based on one or more of the following: (1) sensitive to disruptions in test performance due to an environmental stimulus (e.g., ship motion); (2) concurrence in the scientific literature that the test measured an identifiable information processing or cognition construct for which a theoretical basis was available; (3) differentiation of brain damaged from normal subjects on the basis of test results; (4) inherent interest to the subject; (5) previous appearance in an established and/or factor analyzed battery; (6) obvious face validity; (7)

availability, cost and other practical considerations (Kennedy, Jones, & Harbeson, 1980). Almost no test met all criteria, but most tests met several.

Evaluation Criteria

Stability. Jones, Kennedy, & Bittner (1981) make the point that most subjects demonstrate improvement with practice for most performance tasks (tapping and time estimation are some identifiable exceptions). Performance typically follows a pattern of negative acceleration (ie., classic learning curve for acquisition) with most change occurring early in practice and less occurring late. In general, as practice continues, a subject's performance usually becomes consistent (ie., remains constant or changes in a linear manner over trials). An obvious consequence of such a pattern is that the obtained point measures for a subject, may differ significantly over time. A second consequence of particular concern is the fact that different subjects may respond differently, rather than uniformly, to repeated exposures of the task. Therefore, the relative standings of subjects on the first measure may not resemble the relative standings on the final measure. Only after relative standings are clearly and consistently established between subjects (ie., asymptotic performance with parallel curves for subjects) can the investigator place confidence in the adequacy of his measure. Such an instrument is said to have "stabilized," and results from a stable test may be more readily interpreted, whereas results from unstable tests are ambiguous (Jones, 1979, 1980b, c). Similarly, Jones suggests that repeated measures studies of environmental influences on performance require stable measures if changes in the treatment (i.e., the environment) are to be meaningfully related to changes in performance. Kennedy, Bittner, & Harbeson (1980) call into serious question most previous environmental repeated measures studies which have not addressed the question of stability. They caution that unstable measures "cannot be used reliably to measure environmental change (or any other) effects." (p.3)

Generally stated, a test is defined as stable when: (1) the group means for successive trials become constant (ie., are level, asymptotic or exhibit constant slope); (2) the between subject variances for successive trials become constant (ie., homogeneity of variance); (3) the correlations between a trial and subsequent trials become constant. This latter criterion of stability has been labelled "differential stability," Jones (1969, 1972). If a task has not stabilized, the correlations among successive trials will very likely show "superdiagonal form" (Jones, 1969). That is, the correlations are greatest between two immediately adjacent trials, with greater separation between trials resulting in progressively smaller correlations. Jones (1979) has summarized the superdiagonal form with the following statement:

$$r_{ij} > r_{jk} \quad \text{and} \quad r_{ik} < r_{jk} \qquad (i < j < k).$$

Examination of an intertrial correlation matrix of an unstabilized task makes the pattern readily apparent. Correlations within rows decrease from left to right and correlations within columns decrease from bottom to top. Therefore, the smallest intertrial correlation would be found in the upper righthand corner of the matrix.

When these correlations cease to change within a row and column, and subsequent rows and columns of the matrix, differential stability has been achieved. Theoretically, correlations among stabilized trials are equal. More detailed reviews and specific procedures for statistically establishing test stability may be found in Jones (1969, 1979, 1980b, c) and Bittner & Carter (1981). Examples of applications in establishing test stability may be examined in Harbeson, Kennedy, & Bittner (1979) and Kennedy, Carter, & Bittner (1980). It is important to note that all three of the indicators must be examined in order to assess test stability. It should be noted that differential stability requires not only that both means and standard deviations become constant but intertrial correlations must be symmetrical (Kennedy, Carter, & Bittner, 1980).

Stabilization Time. It may be necessary to evaluate highly transitory changes in performance when studying the effects of various treatments, drugs or environmental stress. Data collected in such situations must clearly reflect effects on performance due to a specific factor, as opposed to confounded effects, resulting from combined factors. Therefore, in addition to stability per se, "good" performance measures should reach stability "quickly," following short versus long periods of practice without sacrificing metric qualities. Clearly, rapidly stabilizing tasks are prime candidates for inclusion in a final battery. A task under consideration for environmental research must be represented in terms of the number of trials necessary to establish stability and/or the total amount of time necessary to establish stability. One task, Grammatical Reasoning (Baddeley, 1968), is representative of tasks that stabilize quickly. According to Carter, et al. (1981), Grammatical Reasoning can be expected to stabilize within five 60-second trials.

Task Definition. Once differential stability has been achieved, the next requirement advocated for a test is task definition. Task definition is the average reliability of the stabilized task (Jones, 1980b, c). Higher average reliability improves power in repeated measures studies when variances are constant. It is well known that the lower the error within a measure, the greater the likelihood that mean differences will be detected, provided variances are also well behaved. Therefore, tasks with low task definition are insensitive to such differences and are to be avoided. For a detailed review of task definition, the reader is referred to Jones (1979). Since different tasks stabilize at different levels, task definition becomes an important criterion to task selection. However, task definitions for different tests cannot be directly compared without first standardizing tests for test length.

Reliability-Efficiency. Test reliability is known to be influenced by test length (Guilford, 1965). Tests with longer administration times and/or more items enjoy a reliability advantage over shorter tests. Therefore, test length must be equalized before meaningful comparisons can be made. A useful tool for making such relative judgments is the reliability-efficiency (also referenced as "standardized reliability, Kennedy, Carter, & Bittner (1980)) of the test. Reliability-efficiency is obtained by correcting a test to a standard administration base time (we employ three minutes) with the Spearman Brown formula. Reliability-efficiency not only facilitates judgments concerning different tests but also provides a means for comparing the sensitivity of one test with the sensitivity of another test.

Task Sensitivity. Task sensitivity may be conceptualized as a test's ability to discriminate differences between subjects on one testing occasion, or within subjects on repeated testing occasions. If tests are stable,

insensitivity is proportional to the lack of reliability-efficiency. In a repeated measures paradigm, each subject serves as his own control and if between subject differences are present, tests with retest reliabilities below r=.25 can be expected to be insensitive to change. Thus, while high task definition (r>.707) does not guarantee sensitivity, lack of it guarantees insensitivity.

Task Ceiling. Tests may meet all of the previously stated criteria and yet be unsuitable candidates for inclusion in a performance battery. Group variability over trials should not decrease. If variability between individual scores decreases over repeated measures, then tests are likely to possess ceilings. If all individual subjects asymptote at the same or near same levels of performance, then the test is said to have a ceiling or top (Jones, 1980 a). Ceilings are undesirable because there is no discrimination measurable between subjects even though discrimination is expected to be present and because overlearning could make performance quite resistant to the environmental treatment. When subjects perform equally well except for random error, between trial correlations fall to zero. This collapse of nonerror variance has been described as "radical destabilization" by Jones (1980b, c). More detailed reviews of the criteria cited above may be found in Jones (1979, 1980 b, c) and application of the criteria to test selection may be examined in Bittner & Carter (1981); Kennedy, Bittner, & Harbeson (1980).

Micro-Based Testing. Environmental performance testing in exotic environments requires that special attention be applied to the testing system as well as the test battery. Features that recommend micro-based testing systems include capabilities for fully automated test battery administration and data storage, portability of the system, as well as reduced size and weight. Also, lost or misplaced data and uninterpretable responses cease to be a common problem of testing. Automated field testing of performance is not without precedent. Wilkinson & Houghton (1982) have adapted a simple reaction time test, known to be sensitive to environmental influences, to a battery-powered cassette recorder. These researchers concluded that the automated mode facilitated environmental testing while preserving the metric qualities of the test. A micro-based approach would preserve the positive aspects of automated testing while providing for greater versatility and flexibility.

Purpose

Because conversion of paper/pencil performance tasks to a micro-based testing mode may alter the metric qualities of the tasks (Wilkinson & Houghton, 1982), the purpose of the present study was to assess the effects of converting "good" paper-and-pencil performance tasks to the automated testing mode. To accomplish this purpose, a group of subjects was administered the same test in both modes. Performance in each mode was examined and the results from the two modes compared.

METHOD

Subjects

Twenty-three Casper College summer school students were recruited for participation. The subjects were solicited from introductory psychology classes on a voluntary basis in accordance with American Psychological Association principles for research with human subjects (American Psychological Association, 1973). The subjects ranged in age from 18 to 47, were in good physical and mental health, and varied from freshman to senior standing. Seven males and 16 females originally volunteered, with one male and two females attriting the study. For two of the subjects, attrition was determined to be related to personal decisions to withdraw from the academic setting as opposed to terminating study participation. In the remaining case, the subjects' data were withheld from analysis due to inability to comply with standard test directions. Final analyses were based on data from N=20 subjects. Subject motivation to participate was high with 62% of those solicited volunteering and motivation for the research task appeared to be high throughout the experimental sessions.

Materials

Previous research with the PETER model identified 30 paper pencil tests as "good" candidates for performance testing. A summary of these efforts appears in Bittner, Carter, Kennedy, Harbeson, and Krause (1984). Five of the tests were elected for further study as possible candidates in a micro-based testing system. Selection of the five tests was based on the following considerations: (1) conformity to the criteria for good performance tests stated above; (2) potential compatibility with the micro-based testing mode; and (3) indications of representing important and well-differentiated factors. The tests, complete with summarized paper pencil selection criteria may be viewed in Table 1. Two of the tests (Spoke and Aim) were not directly adaptable to the micro-based mode. For this reason, tapping tests using key press were programmed to be studied as comparable micro-based tests. Therefore, selection criteria data for Spoke and Aim have been included. More complete reviews for each test may be found in the evaluation references cited in the last column of the table.

Aiming. The Aiming task (Fleishman & Ellison, 1962) is accomplished by accurately marking a dot within a small oval shaped target. The targets were 2 mm in width and were repeated across the test page at the rate of 1/5 mm. Subjects worked continuously following the target trace. Performance was scored according to the number of targets attempted minus the number of targets missed, equal to the number of hits. Aim has been described as a test of manual dexterity with wrist-finger speed, and fine eye-hand coordination important to task performance (Carter, Kennedy, & Bittner, 1980). According to Bittner, Carter, Kennedy, Harbeson, and Krause (1984), "Aim directly provides for assessment of environmental effects on fine eye-hand coordination and indirectly provides for the separation of such effects from other cognitive measures."

Spoke. The Spoke Test (Bittner, Lundy, Kennedy, & Harbeson, 1982) is a modification of the Trail Making Test (Reitan, 1955). The subjects' task was to accurately make a mark within a circular target. The targets were 1 cm in diameter, 9 cm from a control point and were evenly spaced on 32 imaginary radii eminating from the control point. Subjects accomplished the task by starting from the control point, marking the first target, returning to the control point and proceeding to the following target. The task was repeated as often as

TABLE 1. FIVE PAPER/PENCIL TASKS IDENTIFIED FOR POTENTIAL INCLUSION IN A MICRO-BASED TESTING MODE

TASK	TRIAL X STABILIZES	TRIAL SD STABILIZES	TRIAL R STABILIZES	RELIABILITY EFFICIENCY OF r (a)	EVALUATION REFERENCE
AIM	<2	<2	5	.87	Krause & Wolstad (1983); Fleishman & Ellison (1962)
SPOKE CONTROL C	1	1	1	.95	Bittner, Lundy, Kennedy, & Harbeson (1982)
PATTERN COMPARISON	9	9	9	.93	Shannon, Carter, & Boudreau (in press); Klein & Armitage (1979); Carter & Sbisa (1982)
GRAMMATICAL REASONING	4	1	5	.93	Baddley (1968); Bittner, et al. (1984); Carter, Kennedy, & Bittner (1981)
CODE SUBSTITUTION	8	8	8	.84	Pepper, Kennedy, Bittner & Wilkes (1980); Wechsler (1981)

a. Reliability-Efficiency: Reliability estimate for a 3-minute test--computed using Spearman-Brown Formula (Winer, 1971)

b. Task Sensitivity: ++ = $r \geq .8$; + = $.8 \geq .7$

possible in the allotted time. Performance was scored according to the number of targets attempted minus the number of misses, equal to the number of hits. Spoke is a psychomotor task with visual search as an important factor in performance (Kennedy & Bittner, 1978). According to Bittner, Carter, Kennedy, Harbeson, and Krause (1984), Spoke "directly assesses arm movement speed and indirectly provides for distinction of gross environmental disruptions from disruptions in fine eye-hand coordination and cognition."

Pattern Comparison. The Pattern Comparison (Bittner, Carter, Kennedy, Harbeson, & Krause, 1984; Klein & Armitage, 1979) task was accomplished by the subject examining a pair of dot patterns and determining whether they were similar or different. Patterns were randomly generated with similar and different pairs presented in random order. Performance was scored according to the number of pairs correctly identified. Pattern Comparison has been described as a spatial perception task with spatial ability important to test performance. According to Bittner, Carter, Kennedy, Harbeson, and Krause (1984), Pattern Comparison "assesses an integrative spatial function neuropsychologically associated with the right hemisphere."

Grammatical Reasoning. The Grammatical Reasoning test (Baddeley, 1968; Carter, Kennedy, & Bittner, 1981) involves five grammatical transformations on statements about the relationship between two letters: A and B. The five transformations are: (1) active versus passive construction; (2) true versus false statements; (3) affirmative versus negative phrasing; (4) use of the verb "precedes" versus the verb "follows"; and (5) A versus B mentioned first. There are 32 possible items, and they are arranged in random order. The subjects' task is to respond "True" or "False" depending upon the verity of each statement. Performance was scored according to the number of correct transformations. Grammatical Reasoning is described as measuring "higher mental processes" (Baddeley, 1968) with verbal ability an important factor in test performance (Carter, et al., 1981). According to Bittner et al. (1984), Grammatical reasoning "assesses an analytic cognitive neuropsychological function associated with the left hemisphere.

Code Substitution. The Code Substitution test (Pepper, Kennedy, Bittner, & Wiker, 1980) forms were derived by randomly assigning digits to nine letters. The subjects' task was to repeat the assigned digit code when presented with the test letters (Pepper, et al., 1980). Subjects were not permitted to inspect the letter digit codes prior to starting the test. Performance was scored according to the number of correct substitutions. Code Substitution is described as a visual search type task with encoding and decoding, rote recall and perceptual speed as important factors in performance. According to Bittner, Carter, Kennedy, Harbeson, and Krause (1984), "Code Substitution is a mixed associative memory-perceptual speed task which provides for a traditional assessment of those components not otherwise covered by other measures."

Tapping. Tapping was only presented in the micro-based mode. The task was accomplished by alternately pressing keys on the microprocessor keyboard. The tasks were administered to the preferred hand, nonpreferred hand and to both hands working together. Tapping is a psychomotor skill believed to assess factors common to Aim and Spoke.

Apparatus

Micro-based testing was accomplished with the Essex Corporation Automated Portable Testing System (APTS) (Bittner, Smith, Kennedy, Staley, & Harbeson, 1984), implemented on a NEC PC 8201A microprocessor. The NEC PC 8201A is an eight-bit device configured around an 80C85 microprocessor with 64K internal ROM containing BASIC, TELCOM and a TEXT EDITOR. RAM capacity may be expanded to 96K onboard, divided into three separate 32K banks. An RS-232 interface allows for hook-up to modem, to a CRT or flat panel display, to a "Smart" graphics module, to a printer or to other computer systems. Visual displays are presented on an eight line LCD with 40 characters per line. Memory may be transferred to 32K modules with independent power supplies for storage or mailing. The entire package is light weight (3.8 lbs), compact (110 cm (w) x 40 cm (h) x 130 cm (d) mm) and fully portable with rechargeable nickel cadmium batteries permitting up to four hours of continuous operation.

Pattern Comparison, Grammatical Reasoning and Code Substitution were directly adapted to micro-based testing, but because Aim and Spoke were not readily adaptable, the three tapping tasks (Preferred Hand Tapping, Two Hand Tapping and Nonpreferred Hand Tapping) were substituted. Testing times and orders may be reviewed in Figure 1. The system has been produced expressly for human performance assessment, both in unusual and normal environments. A preliminary field test for compatibility with environmental testing needs has been completed and the system was recommended for continued use. More detailed information regarding the apparatus and software may be found in Essex (1984) and Bittner et al. (1984).

Procedure

Prior to testing, subjects received a brief introduction as to the purpose of the study and were advised regarding the general procedures associated with data collection. Subjects were encouraged to work quickly, accurately and to the best of their abilities. Attempts to raise motivation and reduce test anxiety were made by pointing out that the test batteries were the focus of study, as opposed to the subjects themselves. In our judgment, the subjects were motivated to perform, and not adversely affected by performance anxiety.

Subjects were examined over two consecutive test days, in a modified PETER approach. On each day the subjects first received the paper pencil test battery, followed by the micro-based test battery. Practice (see Table 2)was provided preparatory to the first exposure to each test, in each mode. Subsequently, no further practice or warm-ups were given. Having completed the first session, (a session consisted of the administration of one complete paper pencil battery and one complete micro-based battery) subjects were allowed an intersession rest break (3 minutes) and the process of testing with the paper/pencil battery followed by the micro-based battery was repeated. The subjects were thanked, reminded to reappear the following day for further testing and dismissed. The second day of testing was a simple repetition of first, with the exception of practice and statement of purpose. General instruction, statement of purpose and practice administered during session #1 of the first day lengthened the total laboratory time for a subject by approximately 10 minutes. All subsequent sessions were easily completed within a 20 minute time frame. This approach to testing enabled each subject to be tested four times, with each test mode (ie., AB/AB). A schematic representing test battery order, test order

per battery, test administration time, total time on each task and combined total test time may be viewed in Table 2.

Analysis

The group means, standard deviations and 4 x 4 intersession correlation matrices were calculated for each individual paper/pencil and micro-based test. Group means and standard deviations were examined for evidence of test stabilization and intersession correlations were assessed for evidence of differential stability. Rapid stabilization was expected since theoretically comparable practice was received within both modes. Task definition (magnitude of r after stabilization) was determined and evaluated with regard to test sensitivity, as was the reliability-efficiency of each task (cross session correlations normalized to a 3-minute base). Construct validity was assessed via corrected-for-attenuation correlations between the original and the computerized versions of the tests. Such analyses enabled direct comparison and evaluation of the metric properties, of individual tests, and across test modes.

Factor structure of the batteries was determined by three analyses: 1) computerized tests, analyzed for each session separately, 2) paper and pencil tests, analyzed separately by session and 3) all tests for all sessions in a single analysis. Factor analyses used the principal factors method with squared multiple correlations as communality estimates followed by normalized varimax rotation. Factor extraction was terminated when eigenvalues dropped below unity.

TABLE 2. PAPER/PENCIL AND MICRO-BASED TEST ORDERS, TASK TIMES, TOTAL TIMES ACROSS SESSIONS AND DAYS, AND TOTAL TIMES ACROSS TASKS

	DAY 1					DAY 2				
	SESSION 1			SESSION 2		SESSION 3		SESSION 4		
	Practice Time	Trial Time	Total Time Less Practice	Trial Time	Total Time	Trial Time	Total Time	Trial Time	Total Time	TOTAL TASK TEST TIME LESS PRACTICE
PAPER/PENCIL TEST ORDER										
AIM (2 Trials)	15 (a)	90	180	90	180	90	180	90	180	720
SPOKE (2 Trials)	15	30	60	30	60	30	60	30	60	240
PATTERN COMPARISON	15	105	105	105	105	105	105	105	105	420
GRAMMATICAL REASON	15	105	105	105	105	105	105	105	105	420
CODE SUBSTITUTION	15	90	90	90	90	90	90	90	90	360
TOTAL PAPER/PENCIL TIME	75		540		540		540		540	2,160
MICRO-BASED TEST ORDER										
PREFERRED HAND TAP (2 Trials)	10	10	20	10	20	10	20	10	20	80
PATTERN COMPARISON	15	105	105	105	105	105	105	105	105	420
TWO HAND TAP (2 Trials)	10	10	20	10	20	10	20	10	20	80
GRAMMATICAL REASON	15	105	105	105	105	105	105	105	105	420
NON-PREFERRED HAND TAP (2 Trials)	10	10	20	10	20	10	20	10	20	80
CODE SUBSTITUTION	15	90	90	90	90	90	90	90	90	360
TOTAL MICRO-BASE TIME	75		360		360		360		360	1,440
TOTAL TIME OF BOTH	150		900		900		900		900	3,600 = 60 min. total test

(a) times reported in seconds

RESULTS

General

1. All tests (Figures 1-11), whether paper-and-pencil (P&P) or computerized (Comp), show learning curves of similar form. Note that half of these tests are ability tests, but performances improved over sessions equivalently to the motor-skill tests. Note also that improvement with practice is, on the average, 20% from session 1 to 4. Grammatical Reasoning (Comp) improved most (42%), and Code Substitution (P&P) least (8%). On the average, computerized tests improved about as much (19.3%) as paper-and-pencil (22.4%). Ability tests, as a group, improved slightly more (24%) than motor (17%) tests.

2. The standard deviations for all the tests are essentially constant over sessions. This means that the variances are homogeneous, and it also implies that none of the tests is reaching a ceiling.

3. We have selected response/minute (shown on the right of each figure 1-11) as a common metric to aid in comparison across tests and to depict workload. It may be seen that not all tests take equal time, and we may infer they are also all not of equal difficulty. The range is from a low of 16 response/minute for Grammatical Reasoning to a high of 228/minute for Preferred Hand Tapping.

4. The reliability of the tests is good, particularly considering the small sample size. The range is from r = .53 for Code Substitution (which may include spuriously low scores for one session) to r = .93 for the Spoke test. When corrections for test length are made following the Spearman prophecy formula, the reliability efficiency (Bittner & Carter, 1981) for ALL tests is greater than r = .85.

5. In cases where computer tests and paper-and-pencil tests were directly compared, the paper-and-pencil reliabilities were always higher, but only slightly.

6. The reliabilities for the Motor tests are higher than those for the Cognitive tests, even when adjustments are made for test length.

7. All tests appear to be differentially stable by the last session, but additional sessions will be necessary to be certain.

Specific Tests

Aiming. As expected, the means increase most over the first two sessions, but are quite regular thereafter. Standard deviations are constant. Figure 1 shows 144 responses per minute by session 4, an improvement of 22%. The average correlation for the last three sessions is r=.91, and they appear to be differentially stable. Because this is a 3-minute test, the average correlation efficiency corrected for a 3-minute base is the same as the average correlation.

Spoke. Mean performance evidences a gentle upward trend over sessions and standard deviations are constant. Figure 2 shows an average of 76 responses/minute by session 4, an improvement of 23%. The average of the correlations for the last three sessions is r=.93. Since this test lasted only 60 seconds/day, the reliability efficiency is greater than r>.96, the next highest (to non-preferred hand tapping).

Pattern Comparison. After session 1 with the paper-and-pencil version of this test, performance improved regularly and smoothly on both P&P and Comp versions. The standard deviations are constant and perhaps a little larger with the P&P version. Performance improved more on the P&P (34%) than the computer (11%) version, and the reliabilities of the last 3 sessions were higher for the paper-and-pencil (r=.93) than the Comp (r=.80) version. All performances appear stable by session 3, and reliability efficiencies are r>.90 for both versions.

Grammatical Reasoning. Improvement is gradual over all sessions and similar for P&P and Comp versions. Standard deviations are essentially constant over sessions and comparable for the two forms of this test. Performances improved more on the computerized (42%) than the paper-and-pencil (25%) version. The reliabilities for the last three sessions were slightly higher for the P&P version and all performances appear stable by session 3. Reliability efficiencies were r>.90 for both versions.

Code Substitution. Means increase gradually over four sessions and improvement is better with computer based scores (23%) than P&P. Standard deviations are constant or may increase slightly with the means in the computer version. The average reliability for the last 3 sessions is slightly higher for the P&P test (r=.60) than the computer (r=.53) version. Moreover, session 3 of the computer tests may have some anomalous scores - giving rise to a test retest correlation between session 3 and 4 of r=.32. The remaining correlations for all other combinations of sessions are r=.60 - comparable to the P&P version of the test. Improvement in performance on this test is greater for the computer (23%) than the P&P (8%) version. By session 4, mean response/minute is approximately 38/ minute. Other than the anomaly mentioned above, the scores would appear to be stable by session 4. The reliability efficiency for this test is the lowest of all in this study, but exceeds r=.85 for a 3-minute base, and is higher for the P&P version than for the computer. The factor structure of this test, which is discussed below, implies that further study may be needed.

Tapping.

Preferred Hand. The means increase most from session 1 to 2 and are very regular thereafter. Standard deviations are constant. Performance improved 20% over sessions and performance rate was greater than 225/minute on the average (i.e., almost 6/second). Reliabilities were high (r=.82), and since the test was only 20 seconds long, the reliability efficiency for 3 minutes was very high. Stable reliabilities appeared evident almost from session 1.

Two Hand. Mean scores increased only 9% over sessions and were very regular. Standard deviations were largely constant and correlations were high and quite stable. Response per minute was fastest for this test and exceeded 6/second on the average by session 4.

Non-Preferred Hand. Means increased linearly over sessions. SDs were constant. Response per minute was 200 and improvement was 11%. Retest correlations were higher for this test than any other and were likely to be stable after session 1.

Factor Analyses

Table 3 gives the factor structure obtained from analyses of computerized test versions in each of the four sessions. It indicates the presence of two well-identified factors in the computer battery. (Two-hand tapping was excluded from this analysis because the inclusion of three highly related tapping tests caused the factor to be overdetermined). Factor 1 is clearly a "motor" factor, probably related to response speed; as such, it affects performance on Pattern Comparison and Grammatical Reasoning. Note however that this influence decreases across sessions, and differential speed of input would be unlikely to have any important effect on cognitive tests with extended practice.

Factor 2 is just as definitively a "cognitive" factor, with its importance for various tests changing with practice. The stabilization of structure is an important consideration in the test stability issues discussed earlier. As noted previously, results for Session 3 are anomalous due to a few extreme scores, and must be viewed with caution. Indeed, for this and other analyses, results should be considered indicative rather than conclusive because of the small sample size involved. The clarity of analyses under this constraint is encouraging.

It should be further noted for this, and for the Paper/Pencil analysis in Table 4, that there was an indication of a possible third factor emerging in later sessions that was too poorly defined to be present in all analyses. Its nature is unknown, but it is possibly related to the "automaticity" of responses characteristic of very well practiced skills (cf. Ackerman & Schneider, 1984; Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977). It is also our belief that there is a significant general factor running through both the computer and paper/pencil batteries. This should be explored with larger N, more practice sessions in later studies, and "g" related marker tests.

Table 4 shows a similar two-factor structure for Paper/Pencil versions, athough the results here are not so well defined. There are clearly motor (Factor 1) and cognitive (Factor 2) axes, but the cognitive tests load very heavily on motor speed throughout the sessions. As a later discussion will show, there is reason to believe that these are essentially the same factors as for computerized versions, but the computerized versions appear to stabilize earlier and to be more clearly defined. The instability may be due to the changing nature of Spoke and Aiming, which converge toward becoming the same test with practice (they correlate Ø.9Ø by Session 4, almost at the limit of their reliabilities), and to shifts in Code Substitution from "cognitive" to "motor" and back again as a result of probable strategy changes by subjects.

Table 5 shows the session by session analyses of the combined computerized and paper and pencil tests. Here, Factors 1 and 2 are clearly cognitive factors loading consistently on Grammatical Reasoning and Code Substitution respectively. Factors 3 and 4 are motor factors loading on Tapping and Spoke/Aiming respectively. An interesting aspect of this analysis is the change in factor structure of Pattern Comparison across sessions. Although Pattern Comparison loads heavily on the cognitive factors early in practice, by Session 4 it loads primarily on Factor 4, a motor factor. Obviously it has shifted with practice from a cognitively dominated task to a test mediated by motor coordination, again perhaps an emergence of "automaticity" in the pattern of responses.

TABLE 3

Rotated Factor Matrix for Computerized Tests by Session
(Loadings > 0.50 are in **bold**; loadings < 0.20 are omitted)

TEST	FACTOR 1				FACTOR 2			
	1	2	3	4	1	2	3	4
Patt. Comp.	21	23	**50**	43	**87**	**84**	**72**	**51**
Gramm. Reas.	**77**	**60**	**56**	30		**53**	29	**69**
Code Subst.					**84**	**73**	38	**83**
Tapping								
-Pref Hand	**75**	**76**	**82**	**85**	**51**			22
-Non Pref Hand	**86**	**89**	**65**	**93**				
Eigenvalues	1.95	1.79	1.66	1.87	1.77	1.52	0.79	1.47

Note: Loadings for Session 3 are based on Maximum Likelihood Factors because of failure of the Principal Factors Analysis to converge.

TABLE 4

Rotated Factor Matrix for Paper/Pencil Tests by Session
(Loadings > 0.50 are in **bold**; loadings < 0.20 are omitted)

	FACTOR 1				FACTOR 2			
	SESSION				SESSION			
TEST	1	2	3	4	1	2	3	4
Patt. Comp.	**53**	**76**	**88**	**66**	**73**	49	21	**62**
Gramm. Reas.	**78**		33	**70**		**96**	**86**	
Code Subst.	-25	**74**	56		**86**	-25	-65	**97**
Spoke	**74**	**79**	**85**	**92**				
Aiming	**60**	**91**	**94**	**90**	-26			21
Eigenvalues	1.86	2.60	2.80	2.57	1.37	1.28	1.21	1.37

TABLE 5
Rotated Factor Matrix for Computerized and Paper/Pencil Tests by Session
(Loadings >0.50 are in bold; loadings > 0.20 are omitted)

TEST	V	FACTOR 1 SESSION				FACTOR 2 SESSION				FACTOR 3 SESSION				FACTOR 4 SESSION			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Patt. Comp.	C	69	79	64	35	55	36	43	34			21		41	23	34	78
	P	47	54	51		67	38	45	22			21			61	56	91
Gramm. Reas.	C	83	81	80	93					31	27	27		-30		24	
	P	91	93	95	87					27							41
Code Substit.	C					82	85	88	75								
	P					82	84	83	75			24		-28	35	20	43
Tapping																	
-Pref Hand	C	52		32		35			26	61	95	76	86	28		22	
-Non Pref Hand	C		38				-25			93	60	86	86		53		
Spoke	P	21							28	69		21	46		93	93	71
Aiming	P			26			25	31		33	20	23	46	90	90	84	75

Note: V is test version (P is paper/pencil; C is computerized)

It is important to note that the two versions of the cognitive tests behave in a highly parallel manner with respect to factorial content. In terms of underlying factors, the two versions, while by no means identical, appear to be acceptably interchangeable.

The matrix in Table 6 was obtained by analysis of all tests combined across all sessions. As the table indicates, there are eight (possibly only seven) factors, surprising in view of the different testing modes, the practice effects occurring and the relatively high reliability of the tests. Factor names and interpretations are tentative because of sample size.

Factor 1 is clearly a motor factor. Largest loadings are on Tapping (Non-Preferred Hand), a "novel" data entry task. The Paper/Pencil motor tasks also load moderately in early sessions, with loadings disappearing by the last session, suggesting an acquisition of data entry skill rather than a terminal performance skill. This is tentatively labeled "Speed of Data Entry."

Factor 2 is the Paper/Pencil analogue of Factor 1. It becomes better defined and differentiated from Factor 1 with practice, and runs through the all the cognitive tests, with greatest importance on the Paper/Pencil versions. This appears to be a generalized speed of hand movement or a Paper Motor Factor.

Factor 3 is a strong factor running through both versions of Grammatical Reasoning, with significant secondaries on Pattern Comparison. This is clearly a cognitive factor but markers are insufficient to label it other than "Grammatical Reasoning."

Factor 4 predominates in the Pattern Comparison tests, but has interesting secondaries throughout almost all the other tests, particularly Aiming. This probably represents the ability to respond both quickly and accurately, a "Controlled Speed Factor."

Factor 5 is for all practical purposes a Code Substitution factor. Code Substitution is a factorially complex test, requiring several factors not shared with other tests (see also Factor 7). Factor 7 is a similar factor, more restricted to the Paper/Pencil version. Both apparently involve unique aspects of template matching, with slightly different manifestations in the two different testing modes.

Factor 6 has primary loadings on Preferred Hand Tapping and some secondaries on early Aiming. It appears to be a relatively straightforward "Motor Speed" secondary, probably reflecting a basic dexterity on well-practiced tasks and/or prior keyboard experience.

As with Factor 5, Factor 7 involves elements of template matching unique to paper tests.

Factor 8 is difficult to interpret. It has loadings on the computer version of Code Substitution and on Two Hand Tapping. It may be due to the divergence of Paper/Pencil and computer versions of Code Substitution across practice, or it may be an error factor. Given the small sample size, it may be best ignored.

TABLE 6
Rotated Factor Matrix for Paper/Pencil and Computerized Tests
Combined Across Four Testing Sessions
(Loadings > .50 are in **bold**; loadings < .20 are omitted)

TEST	V	S	FACTOR								h^2
			1	2	3	4	5	6	7	8	
Patt. Comp.	C	1			35	**63**	48	28	-22	20	93
		2			**52**	**70**				23	86
		3		29	34	**78**	28				92
		4		32	34	**77**					89
	P	1		22		**73**		22	25		77
		2	33	30	26	**78**					95
		3	27	45	20	**79**					97
		4	20	47		**78**			21		94
Gramm. Reas.	C	1	31		83						85
		2			**78**	26		27			81
		3		33	**88**			20			97
		4			**93**		26				95
	P	1	29		**76**	22		27			90
		2			**87**	27					91
		3			**88**						91
		4		36	83	20					93
Code Substit.	C	1				32	**76**			-22	81
		2					**77**		31	20	79
		3					48			**70**	83
		4					83	29			82
	P	1			-29		38	37	**64**	23	90
		2		34		43	30		**61**	27	89
		3		27			**60**		**63**		91
		4				42	30		**74**		90
Tapping	C	1	55	36	33	25	28	46			90
(Pref. Hand)		2	40					**84**			89
		3	49	24		25		**68**			89
		4	65					**66**			86
(Two Hand)		1	**64**	38		35		28		29	87
		2	**78**	24		20				48	97
		3	**80**							37	83
		4	79	31				-29		28	92
(Non-Pref. Hand)		1	79	32							84
		2	**87**	25	23						96
		3	**90**		25						94
		4	87	30							93
Spoke	P	1	26	**88**		26					94
		2	36	**89**							96
		3	27	**91**							95
		4		**94**							95
Aiming	P	1	**50**			25		44			95
		2	32	**75**		45		22			95
		3	23	**83**		35					93
		4		**86**		24		25			90
Eigenvalues			7.7	7.3	7.2	6.3	3.6	2.9	2.7	1.8	

Note: V is test version (P -- Paper and Pencil; C -- Computerized)
S is Session Number

Considerable evidence is available throughout these analyses that the two different versions of the batteries represent, test for test, essentially the same skills. With the exception of the motor factors and some aspects of Code Substitution, the tests appear to be sufficiently alike to be substituted for another. For the motor factors there are both common elements and some significant mode-specific characteristics. Whether one or the other version of motor tests is "superior" is a question to be addressed on the basis of sensitivity and ease of administration and scoring. Code Substitution is multifactorial, and its nature seems to change with practice. Further evaluations after extended practice sessions may be valuable in clarifying its basic structure.

DISCUSSION

We administered eleven tests. They may be dichotomized as: cognitive (6) and motor (5) tests, or paper-and-pencil (5) and computer (6) based. All tests were administered over 4 sessions to 21 subjects. Our objectives were to compare the metric properties of the newly computerized tests with what were intended as comparable paper-and-pencil tests. The paper-and-pencil tests were selected on purpose because they had been established previously as excellent tests within the PETER program (Bittner, Carter, Kennedy, Harbeson, & Kraus, 1984). Specifically, they were expected to possess reliability efficiencies greater (we hoped far greater) than r=.70% for three minutes and they were to stabilize quickly. Less evidence was available for their factorial uniqueness, being based on expert opinion (cf., Bittner, Carter, Kennedy, Harbeson, & Krause, 1984), but we were prepared to study these issues.

In general, Figures 1-11 show that the tests were well behaved over sessions and the means and variances appear stable after one or two trials. Provocative tests of differential stability were not performed as in previous studies (cf., Jones, Kennedy, & Bittner, 1981; Bittner & Carter, 1981; Jones, 1979, 1980 a; Bittner, Carter, Kennedy, Harbeson, & Krause, 1984). This omission will be remedied in a follow-on study where more tests, subjects and sessions will be examined. The present study was designed as a pilot effort to probe the feasibility of the NEC PC 8201A as a field data collection unit. As our first attempt, we were prepared for apparatus malfunction, data loss, etc. It is of more than passing interest and, indeed, should be a reported outcome of this experiment that there were no instances of missing data for any reason - a rare occurrence in repeated measures studies.

It should be noted that session 1 for P&P and computer testing took perhaps 45 minutes, although actual testing time was only 9 minutes for P&P tests and 6 minutes for computer tests with about 1 minute practice for each. Moreover, sessions 2, 3 and 4 took less than half that time and were divided evenly between P&P (15 minutes) and computer (15 minutes) testing. Thus, from a practical standpoint, it may only require from 1 to 1.5 hours total testing time to be confident one has achieved stability on these few tests; however, it is likely that once achieved, with only moderate refresher trials, it may be possible to maintain a practiced subject with stable levels of performance with only 6-12 minutes testing daily.

It is perhaps speculation beyond the data, but it appears that the amount of time (in minutes) expended in repeated measures testing may be depicted like any other negatively accelerated learning function and similar to the learning curve we show in Figures 1-11. That is, sessions get shorter with practice. Moreover, with additional sessions, the elapsed time in testing (i.e., session length) probably approaches the aggregate of the minimum amount of time for each test; but, of course, it never reaches that value, any more than the physiological limit of conduction velocity of nerves is reached in reaction time studies. We believe, therefore, that for the practical issue of conducting tests in unusual environments or with possible toxic agents, the experimenter needs to plan for substantial amounts of time from pretraining to stabilization. If this is done, it is possible that one can maintain calibrated subjects who can be trained up, at short notice and with minimal investment in testing time. In our judgment, this probe technique can reveal treatment for performance factor interactions which may have very important diagnostic significance.

REFERENCES

Ackerman, P. L., & Schneider, W. (1984). Individual differences in automatic and controlled information processing (Report No. HARL-ONR-8401). Illinois: Universty of Illinois, Department of Psychology.

American Psychological Association. (1973). Ethical principles in the conduct of research with human participants. Washington, DC: American Psychological Association.

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

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

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

Bittner, A. C., Jr., Smith, M. G., Kennedy, R. S., Staley, C. F., & Harbeson, M. M. (1984). Automated Portable Test (APT) system: Overview and prospects. Proceedings of the 14th Annual Meeting of the Society for Computers in Psychology. San Antonio, TX.

Bittner, A. C., Jr., Carter, R. C., Kennedy, R. S., Harbeson, M. M. (1984). Performance Evaluation Test for Environmental Research (PETER): Evaluation of 112 measures. Proceedings of the 28th Annual Meeting of the Human Factors Society (pp. 11-15). Santa Monica, CA: Human Factors Society.

Carroll, J. B. (1980). Individual difference relations in psychometric and experimental cognitive tasks. Contract No. N00014-77-C-0722. Personnel and Training Research Programs, Psychological Services Division, Office of Naval Research. (NTIS No. AD A086057)

Carter, R. C., Kennedy, R. S., & Bittner, A. C., Jr. (1980). Selection of performance evaluation tests for environmental research. Proceedings of the 24th Annual Meeting of the Human Factors Society (pp. 320-324). Santa Monica, CA: Human Factors Society.

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

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

Christal, R. E. (1981). The need for laboratory research to improve the state-of-the-art in ability testing. Presented at the National Security Industrial Association, DoD Conference on Personnel and Training Factors in Systems Effectiveness, San Diego, CA.

Christal, R. E., Payne, D. L., Weismuller, J., & Anderson, M. S. (1982). Learning abilities measurement program. Paper presented at the Annual Meeting of the Psychonomic Society.

Cramer, D. B. (1982). Countermeasures subgroup summary report (NASA-JSC 18681). In J. L. Homick (Ed.), Space Motion Sickness Workshop Proceedings. Houston, TX: NASA.

Essex Corporation. (1984). Automated portable test system. Orlando, FL. Brochure.

Fleishman, E. A., & Ellison, G. D. (1962). A factor analysis of five manipulative tests. Journal of Applied Psychology, 46, 96-105.

Guilford, J. P. (1965). Fundamental statistics in psychology and education. New York: McGraw-Hill, p. 452.

Harbeson, M. M., Kennedy, R. S., & Bittner, A. C., Jr. (1979). A comparison of the Stroop Test to other tasks for studies of environmental stress. Proceedings of the 12th Annual Meeting of the Human Factors Association of Canada, Bracebridge, Ontario, Canada, 21.1-21.9. Also, Research Report No. NBDL-80R008, NTIS No. AD A111296. New Orleans, LA: Naval Biodynamics Laboratory, 1981; 20-28.

Helmstadter, G. C. (1964). Principles of psychological measurement. New York: Appleton-Century-Crofts, p. 32.

Jones, M. B. (1969). Differential processes in acquisition. In E. A. Bilodeau, & I. M. Bilodeau (Eds.), Principles of skill acquisition. New York: Academic Press.

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

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

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

Jones, M. B. (1979). Stabilization and task definition in a performance test battery (Final Report, Contract No. N0023-79-M-5089). New Orleans, LA: U.S. Naval Aerospace Medical Research Laboratory.

Jones, M. B. (1980a). Extended practice, video games, and pilot training. Unpublished research proposal submitted to Air Force Office of Scientific Research, Bolling AFB, Washington, DC.

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

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

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

Kennedy, R. S., & Bittner, A. C., Jr. (1977). The development of a Navy performance evaluation test for environmental research (PETER). Productivity Enhancement: Personnel Performance Assessment in Navy Systems. San Diego, CA: Naval Personnel Research and Development Center. (NTIS No. AD A056047)

Kennedy, R. S., & Bittner, A. C., Jr. (1978). Progress in the analysis of a Performance Evaluation Test for Environmental Research (PETER). Proceedings of the 22nd Annual Meeting of the Human Factors Society (pp. 29-35). Santa Monica, CA: Human Factors Society.

Kennedy, R. S., Bittner, A. C., Jr., & Harbeson, M. M. (1980). An engineering approach to the standardization of performance evaluation test for environmental research (PETER). Proceedings of the 11th Annual Conference of the Environmental Design Research Association (EDRA). Charleston, SC. Also, Research Report No. 80R004, NTIS No. AD A111180. New Orleans, LA: Naval Biodynamics Laboratory, 1981; 1-7.

Kennedy, R. S., Carter, R. C., & Bittner, A. C., Jr. (1980). A catalogue of performance evaluation tests for environmental research. Proceedings of the 24th Annual Meeting of the Human Factors Society (pp. 334-348). Santa Monica, CA: Human Factors Society. Also, Research Report No. NBDL-80R008, NTIS No. AD A111296. New Orleans, LA: Naval Biodynamics Laboratory, 1981; 8-12.

Kennedy, R. S., Jones, M. B., & Harbeson, M. M. (1980). Assessing productivity and well-being in Navy workplaces. Proceedings of the 13th Annual Meeting of the Human Factors Association of Canada, Ontario, Canada.

Klein, R., & Armitage, R. (1979). Rhythms in human performance: 1-1/2 hour oscillations in cognitive style. Science, 24, 1326-1328.

Krause, M., & Woldstad, J. C. (1983). Massed practice: Does it change the statistical properties of performance tests? (Research Report No. NBDL-83R005). New Orleans, LA: Naval Biodynamics Laboratory. (NTIS No. AD A139338)

O'Donnell, R. D. (1981). Development of a neurophysiological test battery for workload assessment in the U.S Air Force. Proceeding of the International Conference on Cybernetics and Society (pp. 398-402), IEEE, Atlanta, GA..

Payne, D. L. (1982). Establishment of an experimental testing and learning laboratory. Presented at the Fourth International Learning Technology Congress and Exposition of the Society for Applied Learning Technology, Orlando, FL.

Pepper, R. L., Kennedy, R. S., Bittner, A. C., Jr., & Wilker, S. F. (1980). Performance evaluation tests for environmental research (PETER): Code substitution test. Proceedings of the 7th Psychology in the DoD Symposium, USAF Academy.

Reid, G. B., Shingledecker, C. A., Nygren, T. E., & Eggemeier, T. F. (1981). Development of multidimensional subjective measures of workload. Proceedings of the Conference on Cybernetics and Society, IEEE, Atlanta, GA.

Reitan, R. M. (1955). An investigation of the validity of Halstead's measure of biological intelligence. Archives of Neurology and Psychiatry, 73 28-35.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. Psychological Review, 84, 1-66.

Shannon, D. M., Carter, R. C., & Boudreau, Y. A. (In Press). 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 Motivation and Vibration Studies. New Orleans, LA: Naval Biodynamics Laboratory.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. Psychological Review, 84, 127-190.

Thorne, D., Genser, S., Sing, H., & Hegge, F. (1983). Plumbing human performance limits during 72 hours of high task load. The Human as a Limiting Element in Military Systems, DRG Seminar Papers, Defense and Civil Institute of Environmental Medicine, Toronto, Ontario, Canada.

Wechsler, D. (1981). WAIS-R Manual: Wechsler Adult Intelligence Scale - Revised. New York, NY: The Psychological Corporation, 1981.

Wilkinson, R. S., & Houghton, D. (1982). Field test of arousal: A portable reaction timer with data storage. Human Factors, 24, 487-493.

Figure 1. Aiming - P/P

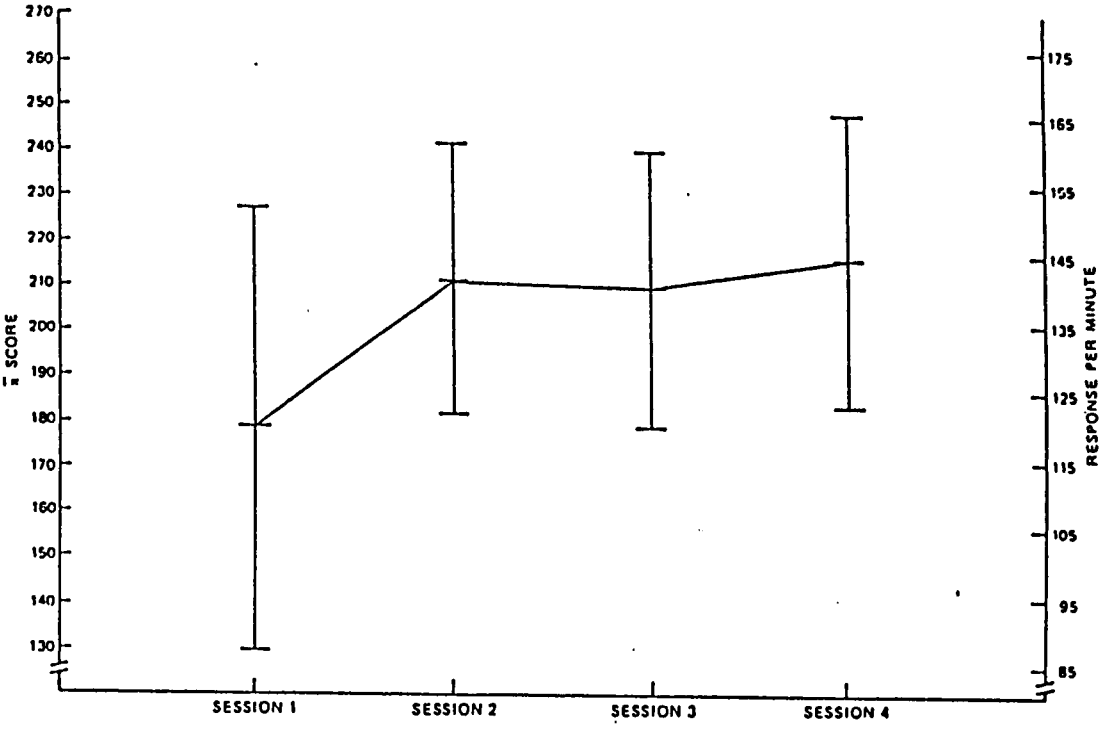

Figure 2. Spoke - P/P

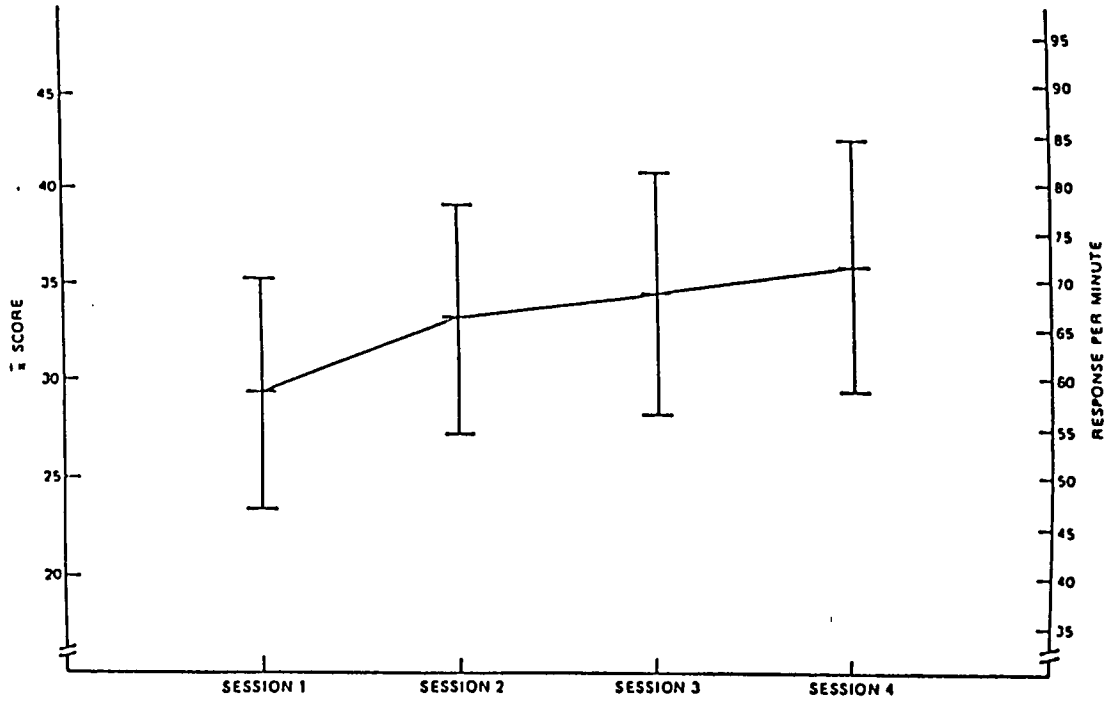

Figure 3. Preferred Hand Tapping - Comp

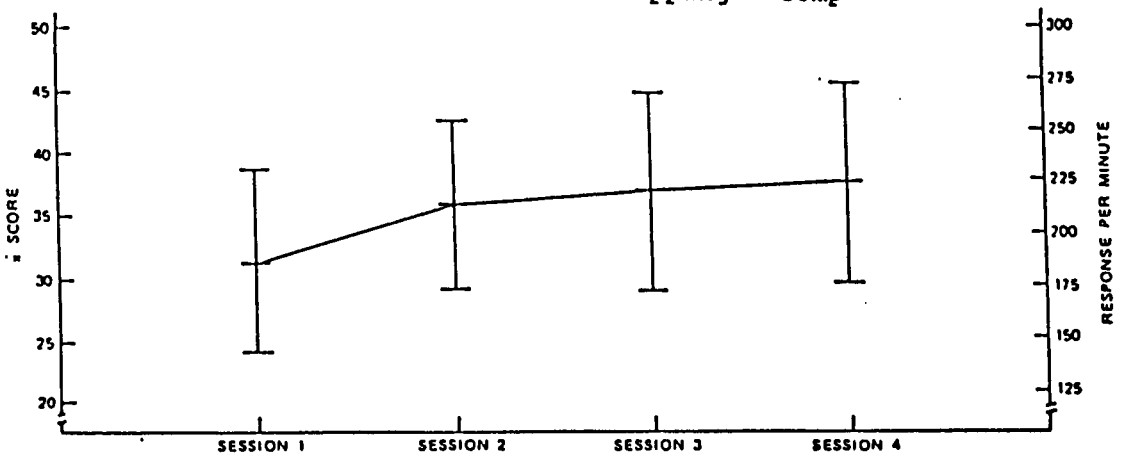

Figure 4. Non-Preferred Hand Tapping - Comp

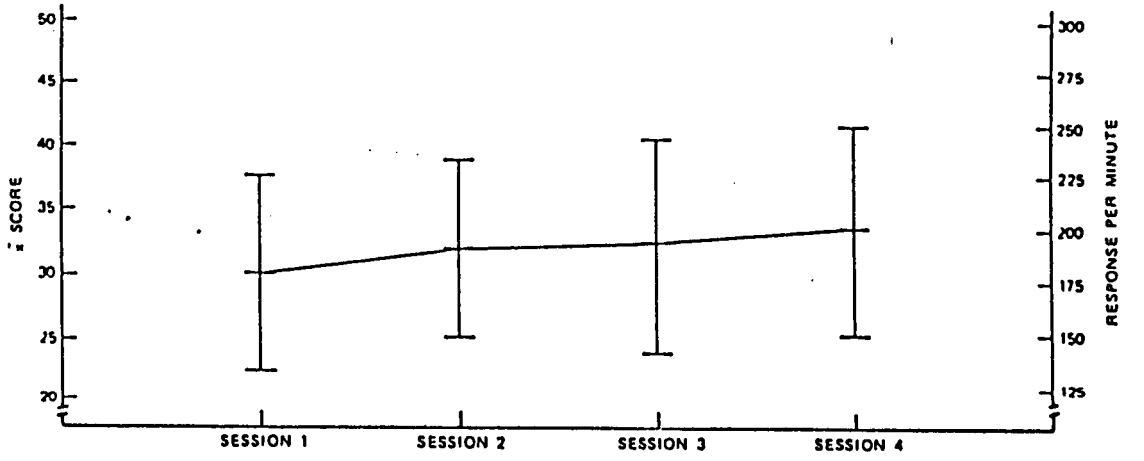

Figure 5. Two-Hand Tapping - Comp

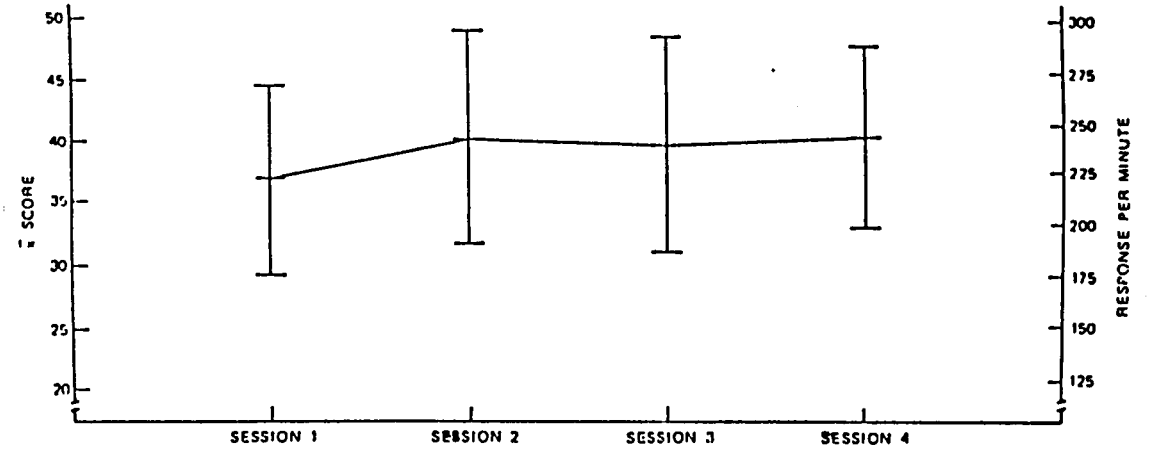

Figure 6. Code Substitution - P/P

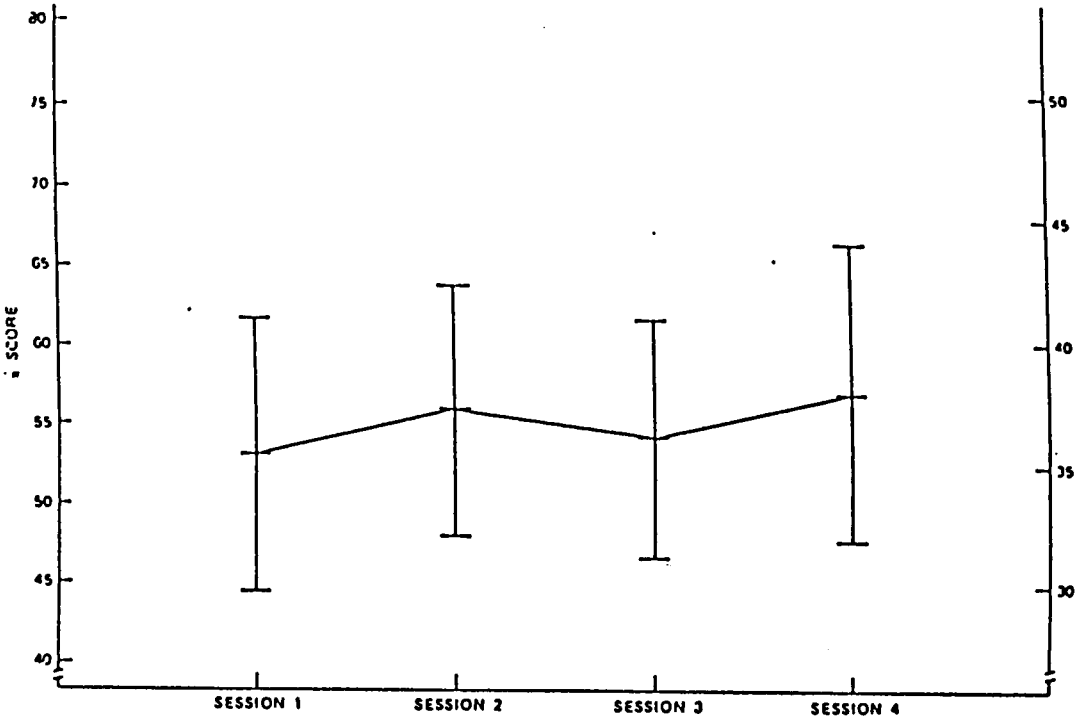

Figure 7. Code Substitution - Comp

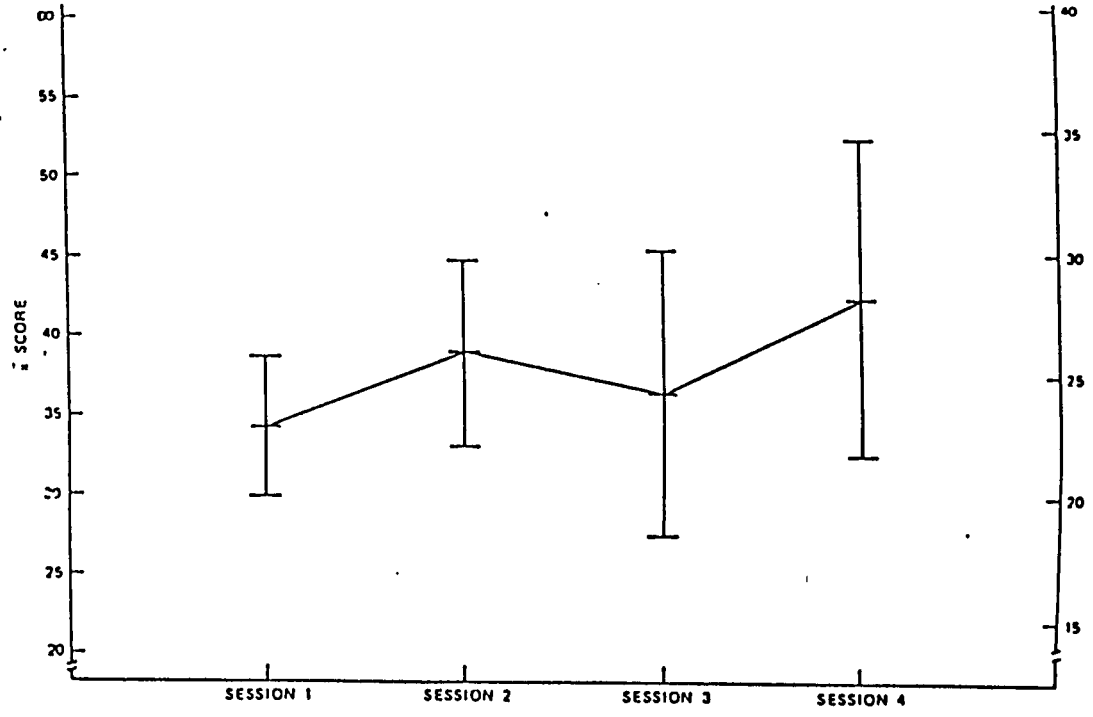

Figure 8. Grammatical Reasoning - P/P

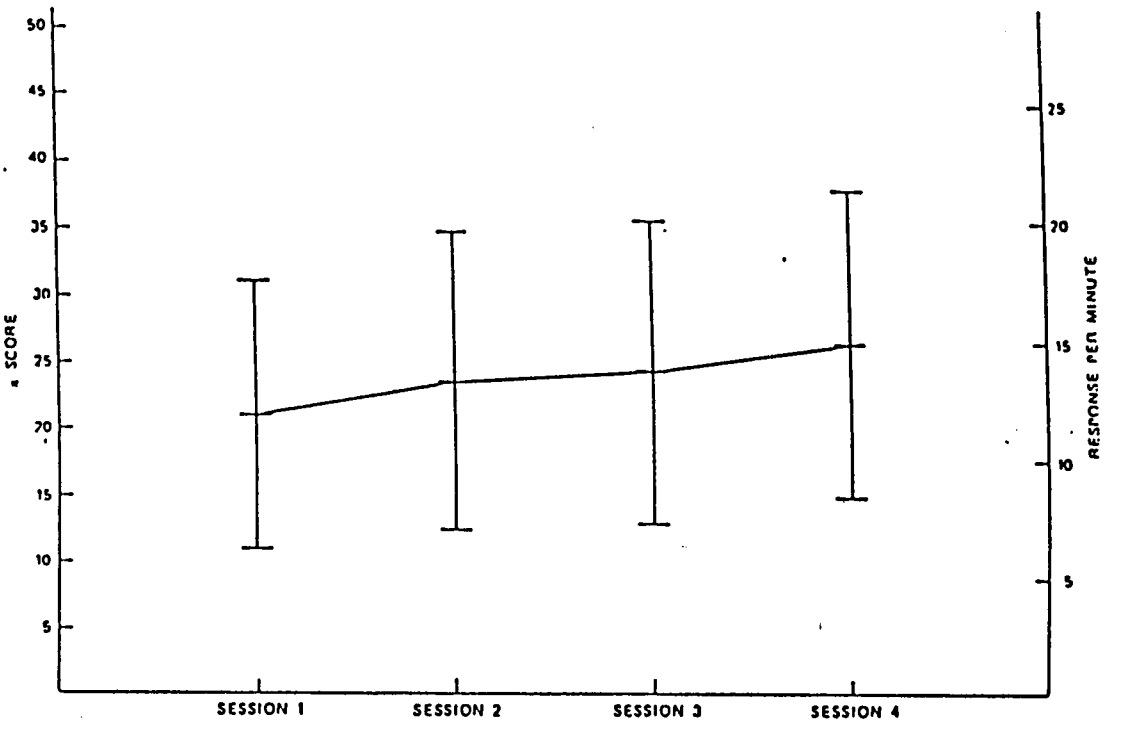

Figure 9. Grammatical Reasoning - Comp

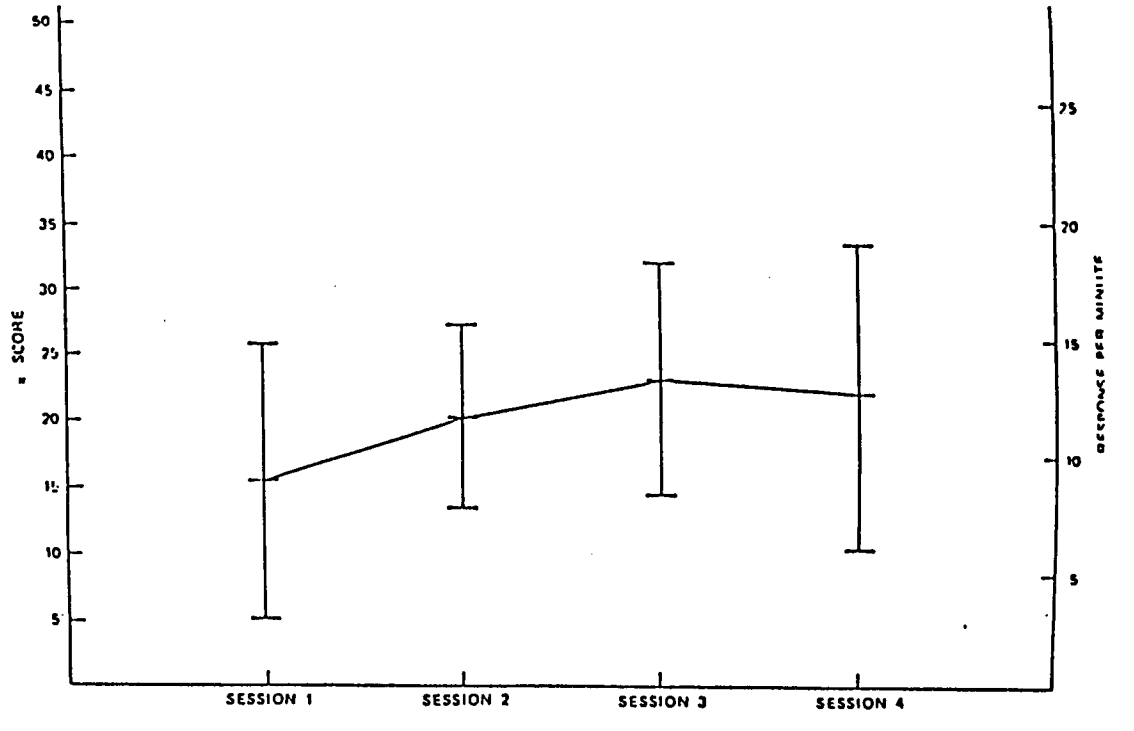

Figure 10. Pattern Comparison - P/P

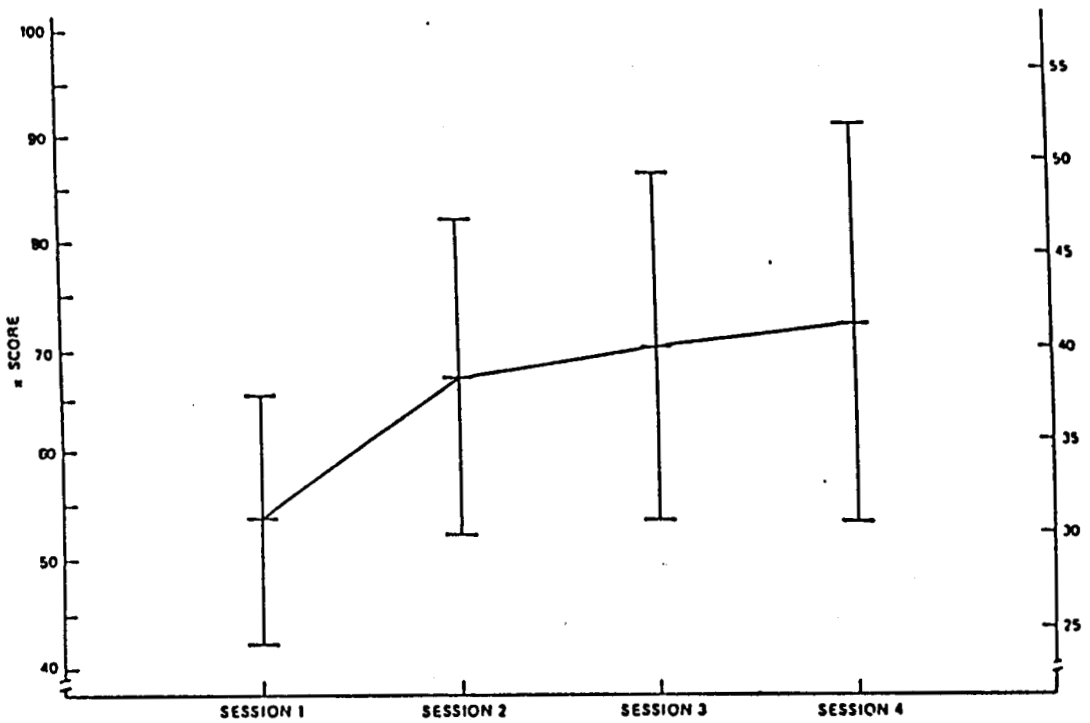

Figure 11. Pattern Comparison - Comp