2

TECHNICAL REPORT EOTR-88-5

USERS MANUAL FOR THE ESSEX AUTOMATED PERFORMANCE TEST SYSTEM [A.P.T.S.]

VOLUME I
USERS MANUAL
APPENDICES A & B

Edited by:

Norman E. Lane
and
Robert S. Kennedy

Essex Corporation
Orlando, Florida

FINAL REPORT
Contract No. DAMD17-85-C-5095

Prepared for:

U. S. Army Aeromedical Research Laboratory
Ft. Rucker, Alabama

May 1988

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

ADA203579

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE

3. DISTRIBUTION / AVAILABILITY OF REPORT
Distribution unlimited;
approved for public release

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
EOTR 88-5

5. MONITORING ORGANIZATION REPORT NUMBER(S)

6a. NAME OF PERFORMING ORGANIZATION
Essex Corporation

6b. OFFICE SYMBOL *(If applicable)*

7a. NAME OF MONITORING ORGANIZATION
U.S. Army Aeromedical Research Laboratory

6c. ADDRESS *(City, State, and ZIP Code)*
1040 Woodcock Road, Suite 227
Orlando, FL 32803

7b. ADDRESS *(City, State, and ZIP Code)*
P.O. Box 577
Fort Rucker, AL 36362-5000

8a. NAME OF FUNDING/SPONSORING ORGANIZATION
US Army Medical Research Acquisition

8b. OFFICE SYMBOL *(If applicable)*
SGRD-RMA-RCD

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
DAMD17-85-C-5095

8c. ADDRESS *(City, State, and ZIP Code)*
Fort Detrick, Frederick, MD 21701-5014

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
	CAL/kmd		

11. TITLE *(Include Security Classification)*
Users Manual for the Essex Automated Performance Test System (A.P.T.S.)

12. PERSONAL AUTHOR(S)
Lane, Norman E., and Kennedy R. S. (editors)

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT *(Year, Month, Day)*	15. PAGE COUNT
FINAL	FROM 4/86 TO 5/88	880516	492

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*

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

This manual describes the characteristics of and provides user information for the Essex Corporation computer-based automated performance test system (APTS). The battery provides integrated software for a menu of 30 performance tests tapping a wide variety of human cognitive and motor funcitons, implemented on a portable computer system suitable for use in both laboratory and field settings for studying the effects of toxic agents and other stressor conditions. From this menu, subsets of tests can be selected using a configuration program to provide a more specialized battery appropriate to the requirements of a particular application. Information on test stability, reliability and factor content is given to assist in decisions about battery composition.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION

22a. NAME OF RESPONSIBLE INDIVIDUAL
Robert L. Stephens, COR

22b. TELEPHONE *(Include Area Code)*
(205) 255-6856

22c. OFFICE SYMBOL
SGRD-UAB

83 APR edition may be used until exhausted.
All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

19. Abstract (continued)

The APTS battery is a consolidation of two independently developed batteries. It contains those 19 of the 25 tests in the UTC-PAB (Unified Tri-Service Cognitive Performance Assessment Battery) that were suitable for implementation on a laptop portable computer intended for field use. It also contains 11 of the tests from the APTS battery developed by Essex for NASA and NSF. The APTS tests have a somewhat longer and more detailed development history, and serve to some extent as anchor or reference points for similar or equivalent PAB tests. Within the combined battery, both sets of tests are implemented through the same menu structure and can be "mixed and matched" as desired for a given application.

The manual gives guidance in selecting, administering and scoring tests from the battery, and reviews the data and studies underlying the battery development. Its main emphasis is on the users of the battery, researchers and echnicians who wish to examine changes in human performance across time or as a function of changes in the conditions under which test data are obtained. Manual content presumes that the user has a general acquaintance with the psychometric properties on which tests can be evaluated, and with the purpose and approach of performance testing in research and field studies.

TABLE OF CONTENTS
VOLUME I

VOLUME II

FOREWORD

The effort described in this document was performed for the U. S. Army Aeromedical Research Laboratory and the Army Medical Research and Development Command under Phase I and Phase II SBIR (Small Business Innovative Research) Programs. The objective was to produce a portable performance test system suitable for use in field studies of the many stressor variables which can impact soldier performance in the operational setting, and to configure the system in such a way that it offered maximum flexibility to the user to tailor a battery to the needs of a specific application. The product of this effort is a menu of 30 tests implemented on a laptop computer, with a configuration program which can be used to select special purpose batteries based on the reliability and factor content of the individual tests in the menu.

ACKNOWLEDGEMENTS

The work reported here represents the cumulative effort of many people over a number of years of literature review, programming, data collection and data analysis, and the concept has benefited from the interlace of ideas from each of these participants. In addition to the editors, each of the following Essex employees and subcontractors has made a substantive contribution to the development and documentation of the battery:

Dennis Baltzley
Judith Chronister
Leilani deSaram
Dr. William Dunlap
J. H. Johnson
Beverly Jones
Dr. Marshall Jones
Lois-Ann Kuntz
Michael Lusito
Mary Kay Osteen
Kevin Schnitzius
Martin Smith
Patrick Smith
Robert Tabler
Dr. Janet Turnage
Robert Wilkes

From the U. S. Army Aeromedical Research Laboratory, Dr. Michael Sanders first recognized the importance of combining existing batteries within a common framework, and emphasized the elements of portability and applicability to field studies. After the departure of Dr. Sanders, Mr. Robert Stephens continued the guidance and shaping of the effort. His contributions are likewise appreciated.

CHAPTER I -- OVERVIEW

This manual describes the characteristics of and provides user information for the battery of tests developed for the U.S. Army Aeromedical Research Laboratory (USAARL). The battery was developed to provide a menu of performance tests tapping the widest possible variety of human cognitive and motor functions, implemented on a portable computer system suitable for use in both laboratory and field settings for studying the effects of toxic agents and other stressors.

The manual gives guidance in selecting, administering and scoring tests from the battery, and reviews the data and studies underlying the development of the battery. Its main emphasis is on the users of the battery, the scientists, researchers and technicians who wish to examine changes in human performance across time or as a function of changes in the conditions under which test data are obtained. Because of this user orientation, the manual inverts the usual order of a research report. The following sections present first the "how to" information needed to make decisions about where and how to use the battery, followed by the research background supporting the battery development. Further, the development history of the battery focuses largely on the logical framework within which tests were evaluated, with technical detail outside that framework provided in a series of appendices.

The battery of tests are a consolidation of two independently developed batteries. It contains those tests from the UTC-PAB (Unified Tri-Service Cognitive Performance Assessment Battery) (Englund, Reeves, Shingledecker, Thorne, Wilson & Hegge, 1987) that were suitable for implementation on a laptop portable computer intended for field use (19 tests out of the 25 specified in the PAB). It also contains 11 of the tests from the APTS (Automated Performance Test System) battery developed by Essex for the National Aeronautics and Space Administration, for the National Science Foundation, and for the U.S. Navy. These two batteries share a number of tests in common, although the way in which the tests are implemented varies somewhat between the batteries. The APTS tests have a somewhat longer and more detailed development history, and serve to some extent as anchor or reference points for similar or equivalent PAB tests. Within the combined battery, both sets of tests are implemented through the same menu structure and can be "mixed and matched" as desired for a given application.

The following sections presume that the user has a general acquaintance with the purpose and general approach of performance testing in research and field studies, and with the psychometric properties by which the "goodness" or "badness" of tests can be examined. For example, the manual will present the reliabilities of tests and discuss the implications of reliability for building a test battery, but will not explore in depth the theoretical underpinnings of reliability concepts and the ways in which reliability coefficients are determined. The objective is rather to guide the knowledgeable user through the procedure of selecting some subset of the 30 tests which is likely to be most effective in a given testing situation, and to speed up and simplify the processes of test planning, administration, scoring and interpretation.

CHAPTER II -- USING THE TEST BATTERY

1.0 INTRODUCTION TO THE MANUAL

Purpose of the Manual

The principal purpose of this manual is to provide user information about the content and metric characteristics of tests in the battery, and to give systematic procedures for determining a test set, configuring a tailored battery or sub-battery from the menu, and using the computer routines which configure the tests in the tailored battery for computer-managed administration and scoring. The manual also provides sufficient additional information about the development of the battery for users to make decisions about test content and test properties, and to understand the process by which tests were judged suitable for inclusion in the battery.

The manual is intended to be used in conjunction with the battery software. There are two main components to the software: The computer tests (a total of 30), and a configuration or setup program which demonstrates each test in the menu and prompts the user to indicate a) the tests to be included in an applications battery, b) the extent of practice time on each test selected, and c) the length of time for each test presentation during a testing period.

Purpose of the Battery

The driving force behind battery development was the requirement to examine changes in a soldier's capability to perform in field settings that might result from one or more environmental, physiological, chemical or psychological "stressor" conditions.

The concept of stressor conditions -- The ability of soldiers to perform tasks in military settings can be affected (usually degraded) by a wide variety of environments and agents. Even well-learned tasks will show performance decrements whenever significant changes occur in a) conditions under which the task is performed (temperature, altitude, visual restrictions, motion, vibration, gravity), b) in the physical status of the operator (fatigue, sleep loss, illness), and/or c) in the biochemical status of the operator (drugs, alcohol, toxic agents and countermeasures, medicines, dehydration, nutritional changes). One of the primary applications of performance test batteries is to study the ability of subjects to sustain performance under such conditions, and particularly to determine the "dosage" effects of the stressor variables involved (the level of stressor at which important performance decrements begin, and the time course of performance changes over continuing stressor exposures.

Constraints on battery development -- These three requirements - field use in addition to laboratory use, application in stress-related conditions, and performance changes resulting from such conditions - serve as primary drivers and "specifications" in shaping the battery, in determining which tests will be included and how they will be implemented.

a) There is a limit on the utility of studies in a laboratory setting for studying stressor effects. It is often difficult, for example, to replicate the combined stressor conditions characteristic of actual operating environments. While laboratory studies can help in "bounding" the problem and in designing field experiments, at some point the battery must go where the "subjects" are.

This need for field use implies a number of constraints on the characteristics of the battery. The first constraint is portability; the test battery must be able to collect data under operational conditions, and tests must thus be usable on battery-operated portable computers or other special devices. Also, for maximum generalization of results, tests used in the laboratory should be the same tests, implemented on the same devices, as those used in the field. Second, soldiers are only available as subjects for limited time periods; the battery must thus require minimum time for practice. The tests must allow soldiers to become familiar with test instructions and to achieve their actual level of performance in only a few practice trials. This is particularly important because tests which require excessive administration time create scheduling difficulties and will interfere with ongoing field operations, causing gaps in the data and seriously reducing the statistical power of field studies. Third, the tests should be free of floor and ceiling effects so that a wide range of ability levels may be studied.

b) The intended application of the battery for studying stressor effects likewise imposes constraints on test characteristics. Because one is usually interested in *changes* in performance that are produced by stressor conditions, it is necessary to repeat the tests several times, to establish a baseline and to examine effects as stressor conditions are varied. This requires that tests be *suitable for repeated measures administration*. Not all types of performance tests can be used in repeated measures designs. For some, the scores are inherently unstable, i.e., scores on successive administrations will never be highly correlated, and the tests will be statistically unreliable. When this is the case, comparison of test scores obtained under stress to each other and to baseline is invalid, since successive scores *do not measure the same thing*. For other tests, the practice trials required for scores on consecutive trials to become correlated may require so much time that it is *impractical* to use the tests under field conditions.

c) Further, when we wish to use tests to assess the degree to which performance is affected by stressor variables, we must have tests which are known to be appropriately sensitive to the widest possible variety of different stressors. By sensitivity, we mean that a test will in general show changes at an intensity of stressor conditions that is comparable to or slightly lower than that likely to be encountered under operational conditions. This is a crucial characteristic of sensitivity; an insensitive test may not show stressor effects until the level of the stressor is so severe as to present a risk of damaging subjects or causing them to abandon the exercise. In addition, when a test for which sensitivity has not been demonstrated is used in a study, a nonsignificant outcome cannot be interpreted, since it cannot be determined if the stressor actually had no effect or if the test variable was simply to insensitive to detect the effect if it were present.

The Concept and Uses of a "Test Menu"

Thirty tests are obviously far more than would ever be practical to use in any study. An extensive body of research suggests that four to eight tests, rarely more than six, are sufficient for examining the effects of virtually any stressor. Although stressor effects on performance appear to be much the same across stressors, performance tends to decrease (with the exception of cold and some drugs) on all tests and to decrease more under greater stressor "dosages"), there may nonetheless be subtle differences in the patterns of test decrements. "Cognitive" tests may drop off earlier than "motor," or the converse; there may be shifts in strategy (e.g., emphasize accuracy over speed); different stressors may interact with modes of stimulus presentation or response. Thus the best "package" of four-to-eight tests for one stressor may be different from the best package for another type of stressor. The "menu" approach used in this battery allows for a wide choice of different tests, and for the convenient construction of smaller batteries tailored to be sensitive to the anticipated effects of the stressors being studied.

Although the number of tests available (30 plus variants) seems relatively large, it should be noted that the tests taken together tap only a limited number of dimensions. Factor analyses indicate that the 30 tests contain no more than five, and possibly as few as three factors, and that most (80% to 90%) of the reliable variance in the battery is present in the first three dimensions. (The "exact" dimensionality of the battery depends to some extent on how a factor is defined and how "important" a factor should be before it is considered "real". There is also a tendency for the factor pattern to change as practice on the tests continues.) Because the number of factors is so small relative to the number of tests, using more than six to eight selected tests adds very little to the information obtained, while materially complicating administration of the battery. A later section gives more detail on the structure oí the battery, an important consideration in using the test menu.

2.0 INTRODUCTION TO THE TESTS

The Test Menu

The 30 tests available in the battery are identified in Table 1. (There are actually 33 tests, with the variants in Tapping, Reaction Time, and Visual/Auditory Counting). Tests labelled as APTS are from the Automated Performance Test System, developed and analyzed with support of the agencies described earlier. Those labelled as PAB are from the UTC Performance Assessment Battery. There is considerable overlap between APTS and PAB with respect to test names. Several of these tests share a common "heritage" in their origins; in most cases, however, the implementation of the tests and the instructions to the subject differ between the two versions.

Description of the Tests

The 19 tests of the PAB are described in detail in Englund, et al. (1987), along with others not implemented in this battery. Brief descriptions of the PAB tests are also included in Appendix A, along with descriptions of the 11 APTS tests. (A total of 21 APTS tests, including six "vision" tests, are completed or under evaluation, but only the 11 indicated were considered sufficiently mature for inclusion in the present menu).

TABLE 1. Tests in the Battery Menu

AUTOMATED PERFORMANCE TEST SYSTEM (APTS)

1. Associative Memory
2. Code Substitution
3. Counting
4. Grammatical Reasoning
5. Manikin
6. Mood Adjective Checklist
7. Number Comparison
8. Pattern Comparison (Simultaneous)
9. Reaction Time
   a. 2 Choice
   b. 4 Choice
10. Sternberg (Short Term Memory)
11. Tapping
    a. Nonpreferred Hand
    b. Preferred Hand
    c. Two-Finger

PERFORMANCE ASSESSMENT BATTERY (PAB)

12. Code Substitution
13. Continuous Recall
14. Grammatical Reasoning
15. Grammatical Reasoning (Symbolic)
16. Item Order
17. Linguistic Processing
18. Manikin
19. Mathematical Processing
20. Matrix Rotation
21. Memory Search
22. Neisser (Visual Scanning)
23. Pattern Comparison (Successive)
24. Pattern Comparison (Simultaneous)
25. Reaction Time (4 Choice)
26. Spatial Processing
27. Stroop
28. Time Wall
29. Vertical Addition
30. Visual Vigilance

The PAB tests in the battery associated with this manual are, with minor exceptions, implemented as described in the PAB documentation. (Tests which require color are either omitted or implemented in monochrome, and some minor changes in instructions were required to eliminate subject inability to understand the task). Also, within the configuration program, there are options for using a system of performance tracking (the Smart System) which verifies subject understanding of instructions and response entry procedures. PAB tests run with the Smart System option have generally higher reliabilities and shorter practice time to stability. The test properties given later in the manual are based largely on data from tests using that option.

Test Properties

There are some critical characteristics about each test that should be considered in the process of deciding which tests to use in a tailored battery. These include a) the number of practice trials (or practice time) required for a test to become "stable," b) its test-retest reliability after stabilization has occurred, and c) its factorial content (what it "measures") both in early trials and in later practice. These properties are in addition to the likely sensitivity of the test to the stressor variable being studied. While information on the first three characteristics is available from a proper test development process, the estimation of test sensitivity to a particular stressor is a much more complex process, and involves some "educated guesswork" based on several different kinds of data and information, most particularly what is known about the stressor itself and about the sensitivity of the tests when used in studies of different stressors. Estimates of stability, reliability and factor structure emerging from the test development process are given below.

Trials to stability -- On the first few trials of practice by an individual on a test, performance is "unstable." Scores on consecutive trials can vary widely, and the ordering of individuals on the test will change, sometimes dramatically, from trial to trial. Once the test is stable, individuals will tend to perform the same way from one trial to the next, means will no longer show large increases with practice, standard deviations will be relatively constant across trials, and, more importantly, the correlations between successive trials for a given test will all be about the same value.

In the study of stressor variables, that is, variables which are expected to create a change in performance, it is absolutely essential that all tests be practiced to stability before any comparison of pre-stressor to post-stressor performance. Prior to the stabilization point, it is not possible to separate the changes resulting from practice from those resulting from stressor effects, and the risk of incorrect inferences is very high. In selecting a battery, preference should be given to tests which stabilize as rapidly as possible so that practice trials can be held to a minimum. Stability is an important concept in test evaluation, and involves examination of means, standard deviations, *and* the magnitude and patterns of intertrial correlations. Evaluation of stability is treated in greater depth in a later section. The second column of Table 2 gives the trial number at which each of the tests in the battery can be considered to be sufficiently stable to examine stressor effects.

Reliability

The higher the reliability of a test, the more one is sure that it is measuring the same thing (construct) from trial to trial. For tests to be used in the study of performance changes, the appropriate reliability coefficient is the "test-retest" correlation obtained from successive administrations of the test, more particularly the average of several different estimates of that coefficient. An unreliable test, e.g., one with intertrial correlations below about .70, may contain too much error of measurement to be useful in repeated measures designs unless it has other overriding properties (unique content, etc.) that warrant its use despite lower reliability. In choosing tests for an application, preference should be given to tests with higher reliabilities. The first column of Table 2 gives reliabilities of the tests in the battery for which sufficient data are available to provide

TABLE 2. Estimated Reliability and Trial of Stability for Tests on the Menu

	Average Reliability Efficiency	Trial of Stability
APTS TESTS		
Associative Memory	.54	5
Code Substitution	.81	2-3
Counting (Audio Counting)	.44	4
Grammatical Reasoning	.86	3
Manikin	.91	3
Mood	NA	NA
Number Comparison	.91	3
Pattern Comparison (Simultaneous)	.85	3
Reaction Time		
a. 2 Choice	.82	3
b. 4 Choice	.83	2
Sternberg (Short Term Memory)	.85	3
Tapping		
Nonpreferred Hand	.98	2-3
Preferred Hand	.98	2-3
Two-Finger	.97	2
PERFORMANCE ASSESSMENT BATTERY		
Code Substitution	.49	7
Continuous Recall	.74	2
Grammatical Reasoning	.71	7
Grammatical Reasoning (Symbolic)	.83	3
Item Order	.31	3
Linguistic Processing	.53	5+
Manikin	.79	3
Mathematical Processing	.64	2
Matrix Rotation	.67	2
Memory Search (Visual-Mixed Set)	.57	2-3
Neisser (Visual Scanning)	.62	3
Pattern Comparison (Successive)	.30	7
Pattern Comparison (Simultaneous)	.46	4
Reaction Time (4 Choice)	.71	5
Spatial Processing	.32	5+
Stroop	NA	NA
Time Wall	.72	2
Vertical Addition	.61	3
Visual Vigilance	NA	NA

NA Indicates either that test was not administered due to hardware constraints or that insufficient data were available for estimation.

an estimate. Note that reliabilities are cast in terms of "reliability-efficiency" estimates. Because some tests require more time than others, and because different time periods were used in different development studies, all estimates have been "normalized" to a three-minute equivalent base. These thus represent the largest reliabilities likely to be encountered in practical applications. A later section will describe ways of adjusting reliability estimates for shorter or longer periods of testing time.

Factorial Content

In tailoring a battery for the study of a particular stressor, it is obviously important to have an indication of what the test measures. The factors on which a test has significant loadings, and the magnitude of those loadings, serve as a guide to understanding test content. There are at least three important factors that consistently recur in various studies of tests in the menu (even in early trials), and a fourth factor that emerges at or around the trial at which most tests are stable. Although factor labelling always involves an element of risk with respect to the "true" content of the factor, a synthesis of factor analysis results across a series of studies suggests the following interpretation.

a) There is in all analyses a factor related to Motor Speed, usually defined by the various Tapping tests, and, in early practice, by the Reaction Time measures as well. This factor also has loadings from other tests for which speed of response execution has an important influence on performance, particularly those for which the "rules" are simple and output is in part dependent on how rapidly responses can be entered.

b) A second factor common to all analyses relates to the facility of the subject with the manipulation of symbolic material using logical rules. This factor, labelled *Symbol Manipulation/Reasoning*, appears to involve a "generalized" ability to reason abstractly through the *application* of rules, rather than the learning or remembering of the rules themselves. While the other factors in the menu are largely *speed*-oriented, and the loadings of the tests tend to change systematically with practice, Symbolic Manipulation/Reasoning tends to show stable loading patterns across trials. It thus may be tapping some inherent capacity related to ability to learn, and not readily changed by practice.

c) A third recurring factor is *Cognitive Processing Speed*. This factor seems to reflect the extent to which defined rules governing *generation of response alternatives* for a particular test have been learned through practice, and can be used progressively more rapidly. To the extent that rules are "mastered," tests loading high on this factor show increases in performance, and the pattern of loadings on Cognitive Processing Speed change systematically with practice. This factor also shows evidence in some studies of heavy loadings on tests with a significant "spatial" manipulation content.

d) A fourth factor emerges in later practice (about trial 4). It is anchored by Reaction Time tests, which become differentiated from the Motor Speed factor after early practice. It appears to involve the speed with which responses can be selected from the generated set of response alternatives, and is thus tentatively labelled Speed of Response Selection.

With the exception of Symbolic Manipulation/Reasoning, which appears to tap a more basic capacity, the factors fit well into a simple conceptual model of information processing and response. Cognitive Speed involves the generation of response alternatives, Speed of Response Selection involves the selection of a response from the set of alternatives, and Motor Speed involves the execution of the selected response. While other interpretations of the results are clearly possible, the interpretation suggested above provides an intuitively appealing framework to which other evidence of factor content can be related.

There are distinct differences in the extent to which the individual tests in the battery load on each of these factors. These differences are a critical aspect of the decision process involved in configuring a battery to be optimally sensitive to a particular stressor. An extended discussion of how particular stressors are likely to affect performance components is beyond the scope of this manual, but relevant information is contained in many of the reports related to battery development (in particular, see Appendix B for a bibliography of studies using various subsets of the battery).

It is difficult to describe in a single table the factor structure(s) of the test battery. The factor patterns obtained from a factor analysis are heavily dependent on the variables included and the size of the correlation matrix analyzed. Likewise, as noted above, there is a well-established tendency for the factorial content of performance tests to change across practice trials. For example, in early practice (particularly the first two trials), most tests involve a component which relates to the ability to understand instructions and to follow directions. This factor decreases in importance for almost all tests as practice continues. Once the subject learns the "rules" for response selection on a test, that test tends to show patterns of loadings which shift systematically toward a factor which assesses the speed with which responses can be generated (i.e., Cognitive Processing Speed).

Given that tests are of limited utility until stabilization has occurred, i.e., there is little change from trial to trial, it is most appropriate to consider the factor structure obtained from stable trials. Table 3 shows the relative importance of factors for each of the tests after most tests have reached

TABLE 3. Factor Structure of Tests in the Menu

	Motor Speed	Symb. Manip./ Reason.	Cog. Proc. Speed	Resp. Select. Speed
APTS TESTS				
Associative Memory (1)				
Code Substitution		++		++
Counting (Audio)(1)				
Grammatical Reasoning		+++		+
Manikin		++		++
Mood (2)				
Number Comparison (1)				
Pattern Comparison (Simul.)		++		++
Reaction Time				
a. 2 Choice				+++
b. 4 Choice				+++
Sternberg (1)				
Tapping				
Nonpreferred Hand	+++			+
Preferred Hand	+++			+
Two-Finger	++			
PERFORMANCE ASSESSMENT BATTERY				
Code Substitution	++		+++	
Continuous Recall		++		+
Grammatical Reasoning		+++	++	
Grammatical Reasoning (Sym.)		+++		+
Item Order	+		++	
Linguistic Processing	++	++	++	
Manikin			+	
Mathematical Processing	+	+	+++	
Matrix Rotation			++	
Memory Search	+	++	++	
Neisser (Visual Scanning)	+		+	
Pattern Comparison (Succ.)	++		+++	
Pattern Comparison (Simul.)	+		+++	
Reaction Time (4 Choice)	+			+++
Spatial Processing	+		+++	
Stroop (2)				
Time Wall				++
Vertical Addition	+	++	+++	
Visual Vigilance (2)				

Notes: (1) following a test indicates insufficient data to estimate loadings; (2) indicates no data collected. +++ (loadings >0.60), ++ (0.40-0.59), + (0.25-0.39) stability (about trial 4 or 5). Since the estimates of loadings and patterns were obtained from a number of different factor analyses over a series of studies, involving differing variable sets and sample sizes, and since a number of these analyses were necessarily based on relatively small numbers of subjects, the loadings are represented in terms of the patterns seen in analyses, rather than in terms of absolute loadings. Loadings are given as High (+++, loading typically greater than .60), Medium (++, loadings between .40 and .59) and Low (+, loadings between .25 and .39). No entry for a variable on a factor indicates an estimated loading below .25. While there is an element of "expert" judgment in such representations of factor patterns, Table 3 likely gives a more accurate picture than that obtained from any one of the several analyses.

3.0 SELECTING TESTS FROM THE MENU

Criteria for Configuring an Applications Battery

The selection of subtests for a battery to be used in a study usually involves a series of explicit tradeoffs. Among these are a) a number of practical constraints cn administration, and b) a critical need to tailor the factorial content of the battery toward those performance components which are most relevant to the purpose of testing and most sensitive to the stressor(s) involved. There are invariably limits on the amount of time subjects can be made available for a single session, and on the number of repeated sessions for which every subject can be reasonably expected to be consecutively available. These constraints will serve as major drivers for deciding how many tests can be in a battery, how much time each test can require, and how many trials to administer.

Likewise, deciding on which particular tests to use in the available time is to a major extent driven by the intended use of the battery and the anticipated effect(s) of the stressor variable. From a "scientific" standpoint, it would be desirable to decide on factorial content first, and then apply the practical constraints to determine how many of the desired tests can be retained in the ultimate battery. In reality, however, the two concerns of content and time cannot truly be addressed separately. An earlier section introduced the concept of "reliability-efficiency," a means of comparing how much useful "information" the individual tests yield when administered for the same amount of time (usually 3 minutes), or conversely, the amount of testing time that must be dedicated to a test to achieve a prespecified reliability (e.g., 0.70). Since it is clear from Table 3 that many tests can be used to tap any given factor, preference should ordinarily be given to those with *higher reliability-efficiency* to achieve more effective use of testing time.

The tradeoffs among such topics as time, content and information efficiency are not conveniently resolved by simple rules or guidance. They involve subject matter knowledge about the effects of specific stressors on humans, about the idiosyncrasies of test content and its changes over practice, about the tests that are likely to be most appropriate for subjects at a particular ability level, and a host of other material well beyond the scope of this manual. The following sections discuss briefly some of the general concerns that should be considered when selecting a battery from the test menu.

Conditions of Data Collection

Testing time available -- It is characteristic of virtually all field studies and most laboratory studies that there are practical limits on the amount of time a single subject will be available in an uninterrupted block of time. A principal determinant of battery content will thus be the total time required to administer a single "run" through the battery. The minimum time required for a single administration of a test selected from the menu varies from less than 30 seconds to around 30 minutes. In addition, the length of time for any given test can be varied using the configuration program, offering considerable control on how much time will be needed for a single "run" through the selected battery.

In general, decisions about the appropriate length for an individual test will be based on information about *time required* for that test *to yield a reliable measure*, and on the degree to which subjects can *maintain sustained concentration or effort* (past about 20 seconds of tapping, for example, muscular fatigue becomes an unintended element of performance). A later section expands on setting test length *after* the battery has been selected. Given below are some guidelines which may be helpful in deciding on how many tests might reasonably be included within a battery.

Test length has a direct effect on test reliability. Most of the tests in the menu (with the exception of Tapping, Reaction Time, Time Wall and the vigilance-based measures) require at least 1.5 minutes of testing time to yield minimally acceptable reliability (two minutes is better), and generally no more than three minutes. Thus the approximate *minimum* time for a single batterv administration (after the orientation or practice trial) can be obtained by multiplying the total number of tests by two and adding about 20 percent to that estimate for transition, administration activities, etc. An 8-test battery would then require at least 20 minutes on the average (although selected tests could lengthen or shorten that time materially), and a good planning estimate would further lengthen that minimum by another 25% to about 25 minutes to allow additional test length. In general, longer is better for both reliability and sensitivity, up to the point at which extraneous factors (fatigue, boredom, loss of concentration) begin to have an impact (about 4 minutes for most tests).

The desirability of estimating the approximate battery length will become more apparent when Table 3 from an earlier section is considered. A well-balanced battery will ordinarily be composed of tests that are representative of all the factors available in the menu, with preference given to those that are most likely to be affected by the stressor. For the four factors in Table 3, a four-test battery would contain one test which is most heavily loaded on each of the factors, an eight-test battery would contain two from each, and so forth, dependent on time available. A six-test battery would "double up" on the factors judged most sensitive to the stressor being studied.

Feasibility of repeated administrations -- There may be a practical limit on the number of times that subjects can return or be made available for repeated trials on the battery. In general, few if any of the tests (Tapping is an exception) are stable on the first trial or two. For the factorial content to be representative of that in later trials, the battery should be administered at least three and preferably four times before the examination of stressor effects begins. Where it is not possible to provide practice for that number of trials, it may be possible to develop a small battery of tests that stabilize very early using the information in Table 2, recognizing that reliability and factorial content will be sacrificed in the process.

There are in addition some tests which may not stabilize for a large number of trials. Although most of the tests in the menu have reasonable properties by the fourth trial, it should be noted that the characteristics given in Tables 2 and 3 are largely for test versions using the "Smart System," a set of algorithms that identify misunderstanding of instructions, random responses, using the wrong keys, and so forth, and significantly accelerate stability (and reliability) by providing additional monitored practice during the orientation trial. It is recommended that the smart system be used on all tests for which it is appropriate, but particularly when early stability is of unusual concern or time constraints are particularly severe.

Degree of experimental control -- The conditions under which tests are administered will vary considerably from study to study. In some, the administrator will be able to spend whatever time is required monitoring the performance of individual subjects and intercepting performance problems that may be unrelated to the purpose of the study. In other settings, particularly in field settings, there is little or no opportunity to monitor individual performance, and the administrator can only "hope" that there are no serious glitches in interpretation of instructions or in willingness to exert effort to perform. Under such conditions, some tests seem to "behave" better than others, that is, they are easier to understand, have less confusing responses, and are in general less susceptible to idiosyncratic behavior. To some extent, identifying these tests involves an element of judgment and some experience with test use and data analysis across a number of applications. APTS Grammatical Reasoning, Pattern Comparison, and Tapping seem to fall consistently into this "dependable" category. In the recommended batteries given in a later section, some preference is given to these more "robust" tests.

Subject motivation -- Although it is important to make sure that subjects have had sufficient practice on the tests before introducing stressor conditions, it should also be recognized that repetitive administrations of the same tests will eventually induce boredom and occasional resistance on the part of subjects. The number of trials for which subjects will maintain maximum effort will vary as a function of such factors as initial motivation, degree of involvement or interest in the study outcomes, and the degree to which subjects perceive that their lack of effort will be detected. When subjects begin to respond randomly, to reverse response patterns, or simply to "coast" through the tests, the tests will begin to "destabilize," reliabilities will drop, overall levels of performance will decrease, and the data will become essentially of no value. Experience with large number of repeated administrations suggests that there is significant danger of such decreased motivation past about seven or eight trials; studies which require trials past that number should consider either reducing the trials through one of the ways discussed above (e.g., using early stabilizing tests), or distributing practice in such a way that repeated administrations are not intensely concentrated in time.

Factorial Content

Table 3 in an earlier section shows the relative factor patterns for the tests in the menu on which data have been collected in one or more of the experiments underlying the battery development. These patterns are extremely significant for selecting a battery from the menu for a specific application. There are two major considerations in using factor content in battery selection -- the nature of the stressor(s) involved and the balancing of the factors or components tapped by the battery.

Nature of the stressor condition -- There are a host of different stressor conditions for which the tests in the menu can provide sensitive batteries. Although the precise effects on performance will differ from one stressor to another, most of the stressors of interest in field or simulated field applications will tend to affect performance through some disruption of the central nervous system (CNS) and its receptor, processor, or effector mechanisms. This suggests that the effects of different stressors will be seen not in the mechanisms that are disrupted, but in the sequence and timing, severity, and "dosage" required to produce performance changes. Such a presumption underlies the idea of a generic battery, applicable across a number of stressor conditions, which allows for comparison of changes for stressors which are not yet well understood to those for which the patterns of disruption are already well established. A later section provides some examples of such generic batteries.

Beyond the concept of generic batteries, there may be other evidence or speculation about stressor effects which would suggest that battery composition should be tailored toward sensitivity to those effects. For example, it is known that well-practiced simple motor tasks are highly resistant to disruption, and performance on such tasks will likely be maintained after tasks with more "cognitive" content have shown distinct decrements. Tests which involve large components of Speed of Response Selection would be somewhat more sensitive to disruption, and those which involve a large component of "processing" to generate response alternatives would be still more sensitive, i.e., they would show decrements at lower levels of the stressor variable. If stressor effects across different levels or "dosages" of the stressor are of interest, it is important to include in the battery some tests which tap each of these "stages" of processing. If, however, the intent is to show that the stressor has the potential for disruption of even the simplest performances, motor and reaction time tests alone may be sufficient to demonstrate the effect. In general, the more that is known about potential stressor effects, the more closely the battery can be tailored for optimum sensitivity.

Balancing the battery -- Some of the tests in the menu, particularly those of the APTS, have been used in a number of stressor studies (hypoxia, altitude, chemotherapy, motion sickness, etc.). (Appendix B provides a list of the documentation from those studies). Experience from these studies suggests that the most useful and generalizable results are obtained from batteries with the greatest factorial richness consistent with available testing time. Even when the effects of a stressor are well understood, comparison of its effects to those of other stressors is facilitated by the use of batteries which contain common tests and which tap as many of the available factors as possible.

There has been considerable discussion within the field of performance testing about the need for "complex" tests. This usually refers to a single test whose performance requires a number of different kinds of abilities, that is, the test itself is factorially complex. Such tests have some serious deficiencies as measures for the study of stressors. They tend to have complex instructions, take a long time to learn, require a great deal of practice before performance begins to "level off," and tend to yield scores which are neither particularly reliable nor diagnostic of the locus of stressor effects, since the scores combine several distinct abilities into a single number. The philosophy of the battery approach versus the single-test approach is to achieve factorial complexity not test-wise, but battery-wise. Thus all the important factors are represented within the battery, but the relative distinctiveness among factors allows for the detection of differential affects across tests and across stressors. The factorial balancing of the battery is an important part of executing that philosophy.

The next section provides some typical batteries "balanced" with respect to factorial content. These are based both on content and on reliability of tests, and both these factors, along with practical concerns (stability, ease of administration, etc.) must be considered. It is important to recognize, however, that a balanced battery means that each of the available factors has neither too many nor too few representative tests. "Overdetermining" a factor can be wasteful of testing time while adding only minimum information; likewise, "underdetermining" a factor obviously omits information that may be important in understanding stressor effects.

In using factor content for battery selection, it should be noted that the factor structure of the tests in the menu is extremely complex. Although a complete exposition of the factor analytic outcomes is beyond the scope of this manual, there are several important findings of the various factor analyses conducted during battery development. These have been discussed earlier, but should be reviewed here. First, there is a systematic shift in the factor composition of the tests from earlier to later trials. The importance of the various factors for a given test (the factor loadings) tends to move as practice continues. In the earlier trials, there are (largely irrelevant) components that reflect the effects of understanding instructions, of general "testwiseness" and of familiarity with the testing media (the computer, keyboard, etc.). These effects tend to decrease in importance with practice, and, as the tests become more stable, the factor patterns tend to become less variant across successive trials. There are also indications in the factor matrices that a greater number of factors are present in later trials, and there is a tendency for communalities to decrease. This indicates that tests are becoming more "test specific" with practice, and share less of their variance with other tests in later trials.

The factor patterns reported in Table 3 are based on trials after tests have stabilized. As such, the table is not representative of the factor composition in the first two or three trials. It is important to remember that the changes in performance across these early trials are almost exclusively the result of practice and test familiarization, and it is not possible to separate these effects from those of any stressor conditions that may be present. It is thus recommended that data from early trials not be compared to outcomes of stressor trials, since the changes in factor composition indicates that something different is being measured during pre-stabilization trials than that measured in later trials.

Some Typical Batteries

Once the approximate time available for a single administration is determined, and the number of tests to be included has been estimated, the next step is to decide on the tests that will be selected for the battery. As Tables 2 and 3 suggest, there are a number of tradeoffs among trial of stability, reliability and factor content, and resolution of these tradeoffs is to some extent idiosyncratic to the test builder's experience and preferences. There are a very large number of different batteries that can be selected from the menu that measure essentially the same mix of abilities. Given below are a series of recommended batteries, ranging from the "core" battery of 5 tests, which can be administered in as little as 8 minutes (10 is better), to a 12-test battery which provides for each factor at least three tests with an important loading on that factor, but requires nearly 30 minutes for a single administration.

CORE BATTERY -- 5 Tests (8-10 Minutes)

Test	Alternate
Nonpreferred Hand Tapping	
APTS 4-Choice Reaction Time	PAB Reaction Time
APTS Code Substitution	PAB Code Substitution
APTS Grammatical Reasoning	PAB Grammatical Reasoning
APTS Pattern Comparison	PAB Patt. Comp.-Simult.

6-TEST BATTERY (11-13 Minutes)

Add APTS Manikin

7-TEST BATTERY (12-14 Minutes)

Add Two-Finger Tapping

8-TEST BATTERY (15-17 Minutes)

Add PAB Math Processing

9-TEST BATTERY (18-20 Minutes)

Add PAB Pattern Comparison-Simultaneous

10-TEST BATTERY (21-23 Minutes)

Add PAB Spatial Processing or PAB Patt. Comp.-Succ.

11-TEST BATTERY (24-26 Minutes)

Add PAB Symbolic Reasoning

12-TEST BATTERY (26-28 Minutes)

Add APTS 2-Choice Reaction Time or PAB Reaction Time

4.0 CONFIGURING THE BATTERY

Through the logical progression of previous manual sections, the user has by this point determined the time available for a single administration, estimated the number of tests to be used, and selected those tests from the menu. Thus far in the decision process, estimates have been made on the presumption that all tests were of the same average length. Before the final test battery software is produced by the configuration program, it is necessary to specify precisely what the exact time and order of presentation for each test will be, and how much time should be provided for the practice or orientation trial. This section provides guidance on selecting test length and practice time, describes the configuration program and its importance in generating the test software, and explains the smart system and its role in achieving most effective use of testing time.

Deciding on Test Length

Beyond the inherent characteristics of the individual tests, the major influence on reliability is the length of the test, the amount of time devoted to presentation of that test in a single administration. It was noted previously that most tests (except for some speed tests) should be run a minimum of 2 minutes, and that longer is better (3 minutes is recommended if time permits). The same test, run for different time periods, will have quite different test-retest reliabilities. Within the "normal" range of times for a test (the 1 to 4 minute range) it is possible to make some quantitative estimates of the effects of adjusting test length using a formula called the "Spearman-Brown" equation (see Winer, 1971, p. 286). This equation, given below, projects the effect on reliability of adding more "items" (for present purpose, items equals time) to a test that are the same as those already included.

$$R_{xx} = n\,(r_{xx}) \,/\, [1 + (n-1)\, r_{xx}]$$

where $\underline{n}$ is the multiplier for test length, $\underline{r}_{xx}$ is the reliability of the shorter test and $\underline{R}_{xx}$ is the reliability of the longer test.

If one knows, either from previous studies or from information such as that in Table 2, what the reliability is for a test of a given duration, the Spearman-Brown can be used to lengthen or shorten the test to achieve some fixed level of reliability judged to be acceptable for a given application. Recall that the reliabilities in Table 2 are obtained from a process called "reliability-efficiency" which projects all tests to an equated or "normalized" length of 3 minutes, the longest recommended time period for normal applications. Throughout the battery development process, a level of about 0.70 has been established (somewhat arbitrarily) as representing "acceptable" reliability for test use. There are, however, many tests in the menu which do not attain this level in a typical two to three minute session. If it is important to use these tests, it may be possible to run them for a longer time, and to shorten others to compensate, or to average scores from several trials to reduce error. In general, it is best simply not to use tests of low reliability, since (as Table 3 shows) there are usually a number of reliable tests available to represent a factor. To offer maximum flexibility in battery development, however, the information in Table 4 is provided.

Table 4, based on the Spearman-Brown formula, provide a means of estimating the effects of adjusting test length. It is cast as a ratio table with the middle column (headed 1.0) as the standard. For example, assuming that 1.0 is the standard (e.g., three minutes), the table is used proportionally. Suppose, for example, that testing time can be increased by 50% for a test with a reliability of 0.64. Then, by looking down column 1.0 to the row with the entry 0.64 and moving along that row to column 1.5, a predicted reliability of 0.73 is obtained.

Deciding on Practice Time

In configuring test batteries throughout the studies underlying this development, the total time allowed for the orientation period has typically been limited to one hour or less. During this time, subjects familiarized themselves with test apparatus, test instructions, and performance requirements, and the test administrator intervened as appropriate to assist table 4 goes here. it will have to printed in compressed print on another printer and inserted here after the page number is determined.

in this familiarization. The lengths of tests during orientation practice were intentionally set at shorter time limits than test trials (usually 30 seconds) to allow subjects to ask questions if they did not understand. These shorter periods of practice for each tests are implemented as defaults within the battery configuration software, but can be varied by the user as desired.

When the Smart System (described below) was implemented to detect subjects who were having problems with the tests (i.e., did not understand instructions), a maximum of five interruptions by the smart system was established to insure that all tests could be presented within the time-frame of the orientation session and to limit subjects' discouragement.

The length and number of practice trials is a complex function of the number of tests and their difficulty level, the characteristics of the subject population, and the clarity of instructions for individual tests. While the default values established for the configuration program represent best judgment about these tradeoffs, the orientation session is crucial to successful administration of later trials, and the structure of the orientation session is extremely sensitive to the overall ability level of the group being tested. It is therefore highly recommended that users pretest practice sessions in brief pilot studies to verify that the default specifications are appropriate for the group on which the study will be performed. If the user determines that more or less time is needed to appropriately orient participants, then the default specification for length and number of practice trials are easily modified using the configuration program.

Using the Configuration Program

The Battery is available on either 5.25" or 3.5" floppy diskette. It is contained on a single diskette, and consists of the following files:

SETUP.EXE	- The battery configuration program
BATTERY.EXE	- The actual test battery
USERS.INF	- Subject information file
ORDER.ORD	- Test parameter file
SUBINFO.DAT	- Current subject information file

SETUP.EXE and BATTERY.EXE are executable files and are initiated by typing their name at the DOS prompt. SETUP is the program which allows tests to be selected from the menu for a battery, and practice time and test time to be specified. BATTERY contains the software for all tests, and the specific tests to be administered in a session are given in ORDER.ORD, which manages the transactions with BATTERY.

The subject information file, while appearing to be an ASCII file, is actually a formatted direct access file, so it should not be opened or viewed with an editor. There is room in the subject information file for 255 subjects. If it does not exist, it will be created by the program. The test parameter file (ORDER.ORD) is an ASCII file, and may be created using the configuration program, or, after some experience with the battery, with an ASCII editor. If ORDER.ORD does not exist, the battery will not run. The format of this file is discussed later. Note that ORDER.ORD must be on the same disk or in the same directory as BATTERY.EXE for the battery to execute.

System requirements -- In addition to the above files, the DOS-supplied ANSI.SYS must be installed on the boot disk. This file is installed automatically in the system at start-up by including the line:

```
DEVICE=ANSI.SYS
```

in the CONFIG.SYS file of the boot disk. Since the BATTERY makes use of special characters, the alternate character set must also be installed. Most versions of DOS supply this with the operating system. On most IBM-compatible systems, the characters are found in a file named 'GRAFTABL', and are loaded by including the line:

```
GRAFTABL
```

in the AUTOEXEC.BAT file. It is easy to tell if the upper 128 characters are loaded or not if the ENTER symbol (a large carriage return or "bent" arrow) is not displayed after pressing ENTER to continue or begin.

Output of the program -- The only "output" of running the configuration program is the creation or modification of the ORDER.ORD file. The configuration program is menu driven, and allows the user to create, inspect or change the ORDER.ORD file. While in the program, it is also possible to invoke a "demonstration" mode, i.e., to select individual tests and to proceed through them to assist in judgments about test configuration. In addition, if the ORDER.ORD file already exists, one can proceed through the selected tests that are listed within the file. The arrow keys are used to highlight the selection. Pressing the ENTER key executes the choice. While in the configuration program, one can usually back-up to a previous page by pressing the ESCape key.

The demonstration capability allows the user to step through the selected battery. There are special keys to press to explore the battery. Pressing the CONTROL-N key will take you to the next test listed in the ORDER.ORD file. If, while "taking" a test, you would like to end and see the score, press the CONTROL-E key. This feature is not available on those tests which do not report a score (e.g. Time Wall, Neisser, Moods, etc.). Keep in mind, however, that it is necessary to get at least one problem correct, otherwise, the "smart" system will take over. When on the page that displays the score, pressing CONTROL-S will display all the statistics collected for that test. These features only work within the SETUP program, and do not work in the BATTERY.

The ORDER.ORD file -- As noted above, the test parameter file, ORDER.ORD, controls the order of test presentation. It requires a number of parameters and special commands to manage execution of the selected battery. Most of those parameters have "default" specifications present in the configuration program. The user may accept tne defaults, or may change them as desired. Keep in mind that not all demand conditions are implemented in each test (e.g., Recall & Linquistic Procesing). This information is displayed after you select a test.

There are four parameters per line: test name, response limitations, practice time and test time, with commas separating each parameter. For example,

PREASON, 15, 1, 3

If this line is found in ORDER.ORD, the battery will execute the PAB Traditional Grammatical Reasoning test. The subject would only be permitted 15 seconds for a response, and would be tested with data collection for three minutes. The first time the subject takes this test, an additional one minute practice session would take place before the actual three minute test. Table 5 lists the test name abbreviations that must be used to identify a test in the file.

The second parameter is the response deadline or time-out parameter. This numeric value, expressed in seconds, is the amount of time a subject will be given to answer a problem. Many of the tests are capable of testing different demand conditions; low, medium, and high. The demand condition desired is conveyed to the test by placing the letter L, M or H immediately after the response time-out parameter. There should be no space between the number and the letter. The only tests which use this format are:

RECALL

MATHP

PSPROC

PLPROC

SREASON

The memory search test (MSERCH) has three variations of presentation; fixed, mixed and varied sets. In addition, there are four different set sizes. Since the response deadline is fixed at 30 seconds for all versions of this task, one designates the number of characters per set (1, 2, 4 or 6) immediately followed by FS, MS or VS for fixed set, mixed set and variable set, respectively.

The Stroop test (STROOP) uses the response time parameter to designate which type of task is being tested. There are three versions of the Stroop: 1) control, 2) interference, and 3) combined. Signify which version to be used by entering 'VS' without the apostrophes followed by the version number (i.e. VS1, VS2 or VS3).

The Mood Adjective Checklist (MOODS) uses the response time parameter to determine how many adjectives to display to the subject. Since there are 50 different adjectives, the range of numbers for this parameter is from one to fifty. Which adjectives will be presented and the order of presentation of adjectives is randomly determined at run time.

The tapping series (TFTAP, P੧TAP, and NPTAP) use the response time parameter to indicate how many tapping trials the subject receives during the test. The type of tapping, (preferred, non-preferred and two-handed), to be performed is determined by the third, or pra^tice time parameter. Instead of a number in this field, the tapping test expects to find a single upper-case character. Legal characters are 'P' for preferred, 'N' for non-preferred, and 'T' for two-handed tapping.

It should be noted that the practice time parameter and the test time parameter can be expressed in minutes or seconds. The battery assumes that any value less than or equal to 15 for these two parameters is minutes. Any value greater than 15 is interpreted to be seconds. The response time value is ALWAYS assumed to be seconds!

There are four more options or special commands available to the user that can be entered into the file using an ASCII editor. If one or more of the following lines are included in the ORDER.ORD file, the battery performs some special functions:

NOPRACTICE
RANDOMIZE
SHOWRESULTS
SMARTSYS=xx

Normally, the first time a subject takes a test, the battery will take the subject through a short practice session of the test before data collection begins. That is, no data is stored in the DATA file for the practice session. The NOPRACTICE parameter disables practice for the subject even if the subject is going through the battery for the first time.

Normally, the battery will take the subject through the selected tests in the exact order listed in the ORDER.ORD file. Including the RANDOMIZE statement in ORDER.ORD will randomly generate a new order of test presentation for each of the included tests each time the subject progresses through the battery.

After a subject has completed a test, the battery will save any data collected and proceed to the next test, without feedback about how well he or she performed on the task. With the inclusion of the SHOWRESULTS statement, the number correct out of the number of problems answered will be displayed for the subject. This is not done, of course, for the Time Wall task and the Mood Adjective Check List.

Functions of the Smart System

The Smart System is entered any time the subject incorrectly answers five or more problems in a row. This feature cannot be defeated. Once the Smart System is entered, the computer continually beeps and displays the subject's score and the number of problems that were incorrectly answered in a row, and asks the subject to contact the experimenter. This feature was added to detect "problem" subjects before they produce potentially unusable data. To get out of this warning loop, press the CONTROL-R key. The subject will then get an opportunity to re-read the instructions and begin the test anew.

The Smart System can also be entered if the subject, after having answered at least 10 problems, has scored at or below the value expressed in the SMARTSYS=xx statement in the ORDER.ORD file. Substitute a numeric value in place of the 'xx' indicating the percentage the subject must surpass. For example, to make sure that all subjects score greater than 50 percent on EACH test in the battery, insert the statement 'SMARTSYS=50' into the ORDER.ORD file. If the subject scores six out of 12, the smart system would be entered, thus assuring that mere chance was not responsible for the subject's scores.

Some of the tests pose a problem to many subjects--so much so, that they enter the smart system three, four and more times in a row! If it is necessary for the subject to proceed to the next test, there is a way to skip a test, but, it is complicated. On most PC compatibles, answer at least one problem correct and then press the ALT key and press '1', '2' and '7' on the numeric keypad, and then release the ALT key. On the Zenith 18x compatibles, simultaneously press the FUNC and ALT keys, and while keeping them depressed type '1', '2' and '7', then release FUNC and ALT. This sequence of keys must be entered in the time provided for a response, so it may take practice to obtain a rapid enough entry.

5.0 ADMINISTERING THE BATTERY

By this point in battery selection, the user has a fully configured battery, implemented in a tailored software configuration, and ready for application. As noted above, it is strongly recommendeu that one or more small pilot studies on the intended subject population be conducted before full-scale data collection is initiated. The procedures in this section apply equally to both preliminary and full-scale studies.

Initial Considerations

Orienting the Administrator -- The administrator should become thoroughly familiar with the selected battery and all aspects of the apparatus prior to data collection, by taking the full battery several times, understanding the instructions from the subjects' viewpoint, and looking for potential "glitches" in the test instructions and sequence. Prior to the orientation session with a subject, the administrator should a) indicate that some of the tests are harder than others and that the subject should expect to have some difficulty with some tests; b) explain that the administrator is available to answer questions and should be contacted immediately should any difficulty arise; and c) advise the subject of the approximate length of the testing session so that the schedule can be adjusted to provide ample time to finish orientation without undue pressure on the subject.

When performing practice trials for familiarization and/or baseline considerations (e.g., to obtain stability for sensitivity testing), the administrator should make sure that the subject is in a normal state of health, and reschedule testing if the subject is overly fatigued, medicated, sick, or otherwise in a condition that would have an adverse impact on test performance.

Hardware Preparation -- Each subject should receive instructions and familiarization with the display and data entry device (usually keyboard) for the testing apparatus.

a) Each subject should be oriented to the visual display and its functions. For example, if using any of the laptop models (e.g., Zenith 181), indicate the location of contrast and brightness adjustments, as well as the tilting screen function. Assist the subject with the adjustments before continuing. Explain that the visual display may be adjusted at any time during testing.

b) Familiarize the subject with the keyboard. Point out important keys and how they are referred to and used in the battery. For example, in the Manikin test, the subject must indicate right or left with the arrow keys; the number keys and the backspace key must be used for the Vertical Addition test. Initial familiarization with the first-time subject is facilitated if the administrator proceeds through the first screens with the subject. Typically, the first screen asks the subject if he/she is a qualified user. If the subject is a new user, the correct entry is "N" for no; if the subject has taken the battery before, type "Y" for yes. The next screen asks the subject to enter his qualifying number. The subject and the experimenter should make a record of the qualifying number entered on the first trial; unless, for example, Social Security Number is used, subjects tend to forget their numbers between administrations.

Using the Smart System

The Smart System serves two purposes. First, it is designed to ensure that each subject understands the tasks and is performing to the best of his/her ability, i.e., not responding randomly, not using the wrong response keys, etc. Second, and perhaps the most useful function, it allows the subject to attain a stable level of performance as quickly as possible (typically in 2 to 3 trials). The Smart System will interrupt a test if the participant scores below a set percentage (usually 60 percent), or answers five problems in a row incorrectly. It will interrupt testing by displaying a message that indicates the percent correct score and/or the number of incorrect answers in a row. For example, "Out of 20 problems you answered 50% correctly or five in a row incorrectly." To reset a test when this message is displayed press the CONTROL key and the "R" key simultaneously. CONTROL R resets the test and allows the subject to review the instructions. It is important to assure the subject each time the Smart System interrupts that this is a common occurrence and does not reflect on their ability. The procedure for restarting a test follows:

1. The FIRST time the Smart System interrupts a test reset the test [CONTROL R] and ask the subject to read the directions carefully.

2. The SECOND time the Smart System interrupts the same test the experimenter should reset the program, read the directions with the subject, and answer questions.

3. The THIRD, FOURTH, AND FIFTH time the Smart System interrupts, reset the test, read the directions with the subject, answer questions, provide examples, and watch as the participant answers the first few problems.

4. If the Smart System interrupts testing for the SIXTH time the subject will undoubtedly feel discouraged. After ascertaining that the subject understands the instructions but simply cannot score high enough to complete the test, he/she must be removed from that test and either dropped from the study or allowed to proceed with the other tests. The procedure for bypassing the test is: (a) press CONTROL R, (b) answer the first problem correctly, (c) press the ALT key (on the Zenith systems, also depress and hold the FUNC key at the same time),

then press the 1, 2 and 7 keys on the numeric keypad successively while holding down ALT (and FUNC on the Zenith). This procedure is deliberately artificial so that a subject is unlikely to discover it by random "playing" with key combinations.

Trouble Shcoting (Commonly encountered problems)

The power is turned off/battery runs down -- (Battery operated systems only) The software is designed to return the subject to the last unfinished test should the computer's power supply fail. After switching the computer back on if it was an accidental shut-off or plugging in the external power source if the batteries were low, have the subject retype his identifying number and resume at the beginning of the test which was left unfinished.

A subject cannot finish the battery -- If the testing design is such that the subject may return to finish the battery, the administrator may simply allow the subject to use the original machine he/she tested on and type in his/her identifier and continue the session. Otherwise, a partial data file for that subject will be taken. For example, during an alcohol study the subject has a relatively small window of opportunity to take the battery (while blood alcohol is high enough) and in the "high alcohol" conditions some subjects may be unable to finish the battery.

Monitoring -- While the battery has been found to be easily self-administered, occasional monitoring is advised. Monitor to ensure that the participant is (a) responding to every problem, (b) pressing the correct keys (i.e., arrow keys for the Manikin), (c) responding with the appropriate hand (i.e., preferred or non-preferred for tapping, (d) adjusting the visual display adequately.

Tests that may require aid

a) Reasoning (Grammatical, Symbolic, PAB and APTS)) - This task requires the participant to comprehend a statement about the order of two letters or symbols and to compare this order with the letters or symbols to the right or below the statement. Many subjects have experienced difficulty comprehending the statement particularly the negative phrasing and statements with word "by". For example, A is preceded by B; or B is not followed by A. The terms "trails" and "precedes" may also need to be defined. Symbolic Reasoning is especially difficult due to the 2.5 second response deadline required in the PAB specifications.

b) Matrix Rotation (PAB) - This task requires the comparison of successive matrices. The participant must enter a ready response by pressing any key on the keyboard after the presentation of the first matrix to signal the next matrix to be presented. The next matrix will not be displayed until the ready response is made. Also, as the directions state, each matrix must be compared to the following matrix. In other words, each matrix should be compared to the next, and a response must be made after every matrix.

c) Time Wall (PAB) - It takes about 3 minutes for Time Wall to calibrate itself. Some computers, such as the Zenith laptops, are equipped with a circuit which, after a period of inactivity, turns off the backlight on the display to conserve battery power. Inspect the computer manual to alter the saving time of the circuit. Due to the delay separating the instructions from the test itself, the subject may require reminding about the instructions.

d) Manikin (APTS AND PAB) - In this test the subject must determine which hand, right or left, is holding the object that matches the object on which the manikin is standing. The manikin may be positioned standing upright facing either toward or away from the subject, or upside down, also facing toward or away from the subject. The manikin's position can be distinguished by characteristics such as facial features and clothing. Some subjects may need to have these characteristic features pointed out and some will have difficulty distinguishing right from left.

6.0 SCORES AND OUTPUT

Layout of the Data Output File

All data that is collected during an exercise is stored in ASCII format in DOS file 'DATA.' This file is found in the current directory on the current logged-on drive. For example, if the battery is initiated from drive A, within directory 'PAB', the data will be found in the file named 'A:\PAB\DATA.' As soon as a subject signs on to the battery, the subject's identification (or qualifying) number, the current date, and the current time are appended to the data file. This 32 character field is defined as follows:

Field Name	Columns
Subject ID	1 - 9
Preferred Hand (R or L)	10
Current Battery Administration	11
Current Test to be Completed	12
Current Month	14 & 15
Current Day of Month	17 & 18
Current Year	20 - 23
Current Hour (Military Time)	25 & 26
Current Minute	28 & 29
Current Second	31 & 32

The subject ID field contains whatever characters the subject entered on the initial sign-on page. The current battery administration (CBA) is an upper case letter of the alphabet. The letter 'A' denotes first administration, 'B' denotes second, etc. with 'Z' indicating administration 26. The current test to be completed (CTC) field is similarly identified. The letter 'A' in the CTC field would indicate that the subject will be starting at the first test defined in

ORDER.ORD, 'B' the second, and so on. (Recall that ORDER.ORD is produced by the configuration program). This information is useful in tracking the subject's progress through the battery. If, for example, the computer "hangs-up" during a test, the subject will be routed to the test that corresponds to the CTC. In this way, the order of testing defined by the experimenter is preserved. All data generated for the subject's session immediately follows this line.

There are four different formats of data contained in file 'DATA', however the first eight fields are common to all tests. Each data field is separated by commas. The common fields are:

Field	Field Name	Columns
1	ABBREVIATED TEST NAME	1 - 7
2	1st PARAMETER IN ORDER.ORD	9 - 13
3	ACTUAL TIME IN TEST (SECONDS)	15 - 17
4	ALLOTTED TEST TIME	19 - 21
5	# EXTRANEOUS RESPONSES	23 - 25
6	# TIMES "SMART" SYSTEM ENTERED	27 - 29
7	TYPE OF FORMAT USED FOR DATA	31 - 33
8	# OF DATA POINTS WHICH FOLLOW	35 - 37

Most of these fields are self-explanatory, therefore, only those fields which require expansion will be discussed. Table 5 lists the abbreviated test names and their associated test. The second field contains the rightmost five characters of the first parameter following the test name in the ORDER.ORD control file. For most of the tests, this field is numeric and

TABLE 5. Test Name Abbreviations

PHTAP	=	Preferred Hand Tapping
NPTAP	=	Non-Preferred Hand Tapping
TFTAP	=	Two-Handed Tapping
PATRNC	=	APTS Pattern Comparison
NUMCMP	=	APTS Number Comparison
MANKIN	=	APTS Manikin
AREASON	=	APTS Grammatical Reasoning
AREACT	=	APTS Reaction Time
STERNB	=	APTS Sternberg
PCODES	=	PAB Code Substitution
MROTAT	=	PAB Matrix Rotation
RECALL	=	PAB Memory Recall
MATHP	=	PAB Mathematical Processing
ITMORD	=	PAB Item Order
MSERCH	=	PAB Memory Search
PATRNS	=	PAB Pattern Comparison Successive
PREASON	=	PAB Grammatical Reasoning
COUNT	=	APTS Complex Counting
PREACT	=	PAB 4 choice Reaction time
SREASON	=	PAB Symbolic Reasoning
PVADD	=	PAB Vertical Addition
TIMEW	=	PAB Time Wall
ACODES	=	APTS Code Substitution
ASSOCM	=	APTS Associative Memory
STROOP	=	PAB Stroop
PPCSIM	=	PAB Pattern Comparison Simultaneous
PMANKN	=	PAB Manikin
MOODS	=	Mood Adjective Checklist
VISVIG	=	PAB Visual Vigilance Task
NEISER	=	PAB Neisser
PSPROC	=	PAB Spatial Processing
PLPROC	=	PAB Linguistic Processing

indicates the maximum response time for a problem presented in the test. However, some cf the tests (e.g., Math Processing, Memory Search) need additional information, and an alphabetic character(s) is appended to the number. Mathematical Processing, for example, determines whether the low, medium, or high demand condition is to be presented by the letters 'L', 'M', and 'H', respectively, following the response time parameter.

The extraneous responses field contains the number of times the subject pressed a key which was not expected. Each test expects certain keypresses as the means of responding to the problem presented. If the subject presses any key other than those expected by the test, this parameter is incremented.

Field seven, type of format used for data, is supplied to enable the experimenter to easily retrieve the data for the test. As stated previously, there are four different formats in the PAB, so this field will contain a numeric value from one to four. Each type of format will be described below.

FORMAT 1. There are seventeen items of data associated with format one. Each datum is a real number within range of -99.999 to 999.999, and each occupies seven character places.

Data Item #	Purpose
1	Number of problems NOT answered (NOTANS) (i.e., response timeout)
2	Number of correct responses (NUMCOR)
3	Number of correct responses less number of incorrect responses (RW) (Rights minus Wrongs)
4	Number of problems answered (NUMANS)
5	Average Response Latency (ARL) for correct responses
6	Standard deviation of ARL for correct responses
7	Highest response latency for correct responses (CORMAX)
8	Lowest response latency for correct responses (CORMIN)
9	ARL for incorrect responses
10	Standard deviation or ARL for incorrect responses
11	Highest response latency for incorrect responses (INCMAX)
12	Lowest response latency for incorrect responses (INCMIN)
13	ARL for ALL responses
14	Standard deviation of overall ARL
15	Median ARL for ALL responses
16	Average lower quartile ARL
17	Average upper quartile ARL

From these data, several derived scores can be calculated. For example:

Number of Incorrect Responses	NUMINC =	NUMANS-NUMCOR
Percent Correct	PERCEN =	(NUMCOR/NUMANS)*100
Total Number of Problems	NPROB =	NUMANS+NOTANS
Overall maximum response time	OVERMAX=	MAX(CORMAX,INCMAX)
Overall minimum response time	OVERMIN=	MIN(CORMIN,INCMIN)
	(only if	INCMIN<>0)

All the times are in seconds, and are attained by using the SECONDS command within BASIC. The SECONDS command uses the system clock of the PC, which interrupts approximately 18.73 times a second. Therefore, the minimum time factor is about five hundredths of a second.

It should be noted that if the subject did not have any incorrect responses, the time values for incorrect responses will be zero.

FORMAT 2 is only used by the tapping tests, and the data consist of the number of alternate keypresses followed by the total number of legitimate keypresses for each trial of the test. Field two of the common field holds the number of trials in each test. Thus, the number of items of data will always be two times the contents of field two of the common data.

FORMAT 3 is used exclusively by the Time Wall test. Like FORMAT 2, there can be a variable amount of data for this test. The Time Wall test calibrates its timing loop for the dropping of the brick each time it is executed due to the speed differences of PC's. To accomplish the 10 second drop of the brick, a delay loop is executed for each horizontal line of the display. The number of times this null loop is executed is saved in the data as data item one. Data item two contains the number of seconds the calibration routine decided was close enough to ten to determine the null loop counter. The remainder of the data items contain pairs of times for each trial. The first number of the pair is the actual time the brick took to fill in the hole. The second number of the pair is the time the subject determined when the brick filled in the hole. The number of trials can be calculated as NTRIALS = ((number of data points)-2) divided by 2, and the format would look like:

Data Item #	Purpose
1	Value used in delay loop
2	Time that calibrate routine exited with
3	Actual time, trial one
4	Subject's time, trial one
	.
	.
	.
(NTRIALS*2)+1	Actual time, trial NTRIALS
(NTRIALS*2)+2	Subject time, trial NTRIALS

FORMAT 4. This format is used by the Mood Adjective Checklist and is a combination of formats one and two. The data items are:

Data Item #	Purpose
1	Overall average response latency
2	Standard deviation of ARL
3	Highest response latency
4	Lowest response latency
5	Median response latency
6	Average lower quartile ARL
7	Average upper quartile ARL

Thereafter, the remaining data items are specially coded integers which contain the adjective number (see Table 6 for list of adjectives and number) and the subject's response to the adjective. To retrieve the adjective number, take the integer division of the data item by eight. To retrieve the subject's response to the adjective, use the integer modulus of the data item with eight. That is,

Adjective Number = INT((data item value)/8), and,

Response Value = (data item value) MOD 8

Response values are coded as:

1 = Definitely Applies

2 = Somewhat Applies

3 = Does Not Apply.

For example, if the data item value were equal 283, using the above formula, the result of the integer division of 283/8 would equal 35, and 283 mod 8 would equal 3. This means that the subject responded does not apply to mood number 35, or OPTIMISTIC.

The number of adjectives responded to can be obtained by subtracting seven from the number of data points found in the common field.

TABLE 6. Adjective Numbers for Mood Adjective Check List

1	=	ACTIVE
2	=	APATHETIC
3	=	APPREHENSIVE
4	=	ATTENTIVE
5	=	BELLIGERENT
6	=	BLUE
7	=	BUSINESS-LIKE
8	=	CHANGEABLE
9	=	CHEERFUL
10	=	CONFIDENT
11	=	CONFUSED
12	=	CO-OPERATIVE
13	=	DECISIVE
14	=	DEPRESSED
15	=	DETACHED
16	=	DISTURBED
17	=	DREAMY
18	=	DULL
19	=	EASYGOING
20	=	ENERGETIC
21	=	ENTERPRISING
22	=	FORCEFUL
23	=	GENIAL
24	=	GOOD-NATURED
25	=	HEADACHE
26	=	HUMOROUS
27	=	IMPATIENT
28	=	IMPULSIVE
29	=	INDUSTRIOUS
30	=	KEYED-UP
31	=	KINDLY
32	=	LEISURELY
33	=	LONELY
34	=	NERVOUS
35	=	OPTIMISTIC
36	=	QUIET
37	=	RELAXED
38	=	SARCASTIC
39	=	SELF-CONFIDENT
40	=	SKEPTICAL
41	=	SLEEPY
42	=	SLUGGISH
43	=	SUBDUED
44	=	TIRED
45	=	TRUSTFUL
46	=	UNEASY
47	=	VIGOROUS
48	=	WILLFUL
49	=	WITHDRAWN
50	=	WORRIED

Scoring

As shown in the previous section, each test in the menu generates a large number of possible scores, most of which are inherently interrelated. Virtually all of the tests in the menu (with a few exceptions such as the Mood Adjective Check List) are administered under fixed time constraints. Further, most of the tests which are not predominantly motor speed or reaction time are designed to be as "easy" as possible for subjects to determine correct answers, that is, by the trial of stability, practiced subjects with an understanding of the instructions should be making only a few errors. These two design aspects, fixed time per test and "easy" tests, tend to make measures of speed and measures of number of correct responses nearly equivalent. This has several implications for choosing a score or scores from the many potential numbers recorded in the data files.

Available scores -- Scores that are directly generated from test responses or that can be derived by simple algebraic manipulation of direct scores fall into one of four general classes. These are: a) Number of correct responses (NC) (this includes number of alternating keystroke pairs in tapping), b) response latency (RL) measures (average latency per response, average latency per correct response), c) percentage of correct responses (PC), and d) correct responses adjusted for guessing (right minus wrong) (RW). There are also a number of variability scores associated with response time that indicate consistency of latencies across items.

Use of all the available scores produced by the tests is impractical (because of the magnitude of data generated), methodologically unsound (because of the risk of chance capitalization) and unnecessary (because most scores yield the same information in different forms). For tests that involve some "cognitive" decision in response selection (Grammatical Reasoning, Manikin, etc.) the total number of items answered in a fixed time period will correlate perfectly with average latency, and, since there should be few errors in practiced subjects, the number correct (NC) will ordinarily correlate with RL as high as their reliabilities allow. Studies underlying this development show, however, that RL measures obtained under time constraints tend to be somewhat *less reliable* than NC measures, and also contain less information, since NC scores are influenced by *both accuracy* of response *and speed* of response.

Percent correct (PC) scores, although in common use for cognitively-oriented tests, have a number of serious deficiencies as performance measures for most of the tests in the menu. Unlike NC, which carries information about both accuracy and speed, PC contains only accuracy information, and is insensitive to response strategies which produce accurate responses rapidly (the usual definition of skilled performance). Only when time is unlimited per item and per test, and tests are unusually difficult, do PC scores add an additional dimension to test information. Further, as subjects become more skilled in later practice, errors disappear and PC goes very high (over 90%), reducing its variance and lowering its reliability, sensitivity and correlation with external variables.

There are variables for which RL or speed-oriented scores are the "natural" measures. For reaction-time tests, for example, NC and PC scores make no sense. Because there are few or no errors, NC will correlate perfectly with RL, and PC will be near unity with little or no variance. Tapping tests, although their metric is cast in terms of number correct, are in essence analogous to time-based or RL measures.

Derived scores -- A number of other scores can be derived from the basic output data, by one or more (usually nonlinear) transforms or by running tests under several different conditions of difficulty. On the Sternberg, for example, slope scores can be obtained by varying the size of the stimulus set. These slope scores, usually based on three or four points per subject, are analogous to correlations with only one or two degrees of freedom (recall that two df are lost in fitting a line), and as such are notoriously unreliable as individual difference measures (Dunlap, Kennedy, Harbeson & Fowlkes, 1988). Likewise, scores obtained by subtraction of quantities from one another (difference scores, ggain or change measures) are also known to be extremely unreliable, and of thus of limited value as performance measures, particularly when sensitivity to stressor effects is a major concern of the study. (See Cronbach and Furby, 1970, and Rogosa, Brandt and Zimowski, 1982, for a thorough discussion of change measurement). Slope, similar measures which involve parameter fitting from the data, and difference or change scores are not recommended for use in any of the tests in the menu.

A second form of derived score can be useful under some circumstances. "Throughput" measures, obtained by dividing the number correct by the average latency of all responses, indicate the "correct answers per unit of time," and can be sensitive to conditions that are not detected by the other measures (Kennedy, Dunlap, Bandaret, Smith & Houston, 1988; Thorne, Genser, Sing & Hegge, 1983). Subjects under sharply degraded performance conditions (high or continuous stress) may shift to a coping strategy of concentrating exclusively on correct responses and ignoring speed. These sometimes abrupt changes in speed-accuracy tradeoff can be identified by decrements in throughput measures, which may drop sharply when only moderate decrements are seen in NC.

Recommended scores -- In general, it is recommended that only one score from each test be used in stressor studies. The scores which appear across several studies to have best reliability, greater sensitivity, and earliest stabilization are the NC scores for "cognitive" tests and RL scores for "speed" tests. The use of NC is recommended for all tests except the following:

a) The three Tapping tests, for which alternate keystrokes is the recommended metric.

b) Reaction Time tests and the Sternberg, for which average reaction time (RL) is recommended.

c) Time Wall, for which no clear single metric is available. The two studies with Time Wall have used the average of differences between actual time of drop and estimated time of drop, with inconsistent results.

d) Mood Adjective Check List. Since this is not a performance test, "scoring" of responses lies in a different domain than other tests in the menu. The output files give considerable information about item responses, and the user is encouraged to derive a scoring system appropriate to a specific application. In some work unrelated to the present development, the Mood has shown sensitivity to a stressor variable (long-term isolation) when there were no detectable performance changes.

Although PC and RL (for tests scored with NC) are in general not recommended as performance scores, they have considerable value as pointers to subject difficulties with instructions, lack of motivation, or other anomalies in the obtained data. Low PC values may indicate lack of understanding of instructions, an overall ability level too low for best use of the battery, or random response strategies. Extremely variable RL scores, for example, can be used to detect apparatus difficulties or subject confusion about appropriate response procedures.

7.0 SUGGESTED ANALYSES

Although the specific analyses performed on the test data will be a function of the design of a particular study, there are a number of standard analysis procedures recommended for data from any application of tests in the menu. Despite the most careful design and administration, there is always some risk of anomalies from atypical behavior by subjects or from problems in a particular device. These anomalies can ordinarily be detected by careful analysis of test characteristics prior to examination of stressor vs. baseline performances. In general, these involve a) initial screening of data distributions for unusual or atypical responses, b) checking for the presence and shape of expected practice effects, c) verifying the presence of test stability, and d) determining the adequacy of test reliability. These analyses are distinct from those involved in comparison of performance under different test conditions (ANOVA, multivariate analysis, etc.), and should be considered as a routine precursor to such statistical tests.

This sequence of recommended analyses is part of the APTS test development paradigm, and has been followed in each of the studies performed for the present development as well as those underlying earlier APTS developments. This section will briefly review the purpose and general approach to these preliminary analyses; the paradigm is discussed at greater length in Chapter

III, and several of the Appendices (particularly E, F, and G) provide detailed examples of each analysis procedure.

Initial Screening of Data

A critical concern in the use of any complex testing system is that subjects clearly understand the instructions and the appropriate responses for each test. Although the Smart System has significantly reduced the problem of instruction misunderstanding, it is still important to examine the descriptive statistics for all tests on all trials, at a minimum the mean, standard deviation, and low and high values, to isolate "impossible" scores and other possible glitches in data collection. It is also desirable to plot the frequency distributions for the same data sets to look for "outliers" or extremely deviant scores that may be the result of some problem with a subject or the testing system, and to inspect individual data across trials to see if patterns emerge for particular subjects. Percent correct and latency measures are extremely useful in identifying unusual response patterns, but, it should be recalled, are not ordinarily recommended for further analyses. Anomalies should be detected and "repaired" before proceeding with further analyses.

Under ideal testing conditions, it would be desirable to perform each of the above analyses on each session's data immediately after that session, before continuing with the study. Although rarely possible, such a refinement avoids many of the risks of unusable data in later trials.

Checking for Practice Effects

One of the most dependable effects in repeated measures testing is that practice leads to improvements in performance. Means increase, variability decreases, and the group performance across trials will show a predictable form. An important part of initial data analysis is to verify that the practice curves for each test are reasonably similar to expectations, i.e., mean performance should not decrease across trials except under very unusual circumstances, and the reasons for any such decreases should be determined. Chief among these reasons (other than introduction of a stressor) is a change in subject motivation (too many trials, boredom, etc.). While individual learning or practice curves are much more variable than those for the group, it may be necessary to plot or otherwise examine individual performance trends to determine the extent to which decreases are general or are caused by a few isolated individuals. Lane (1987, pp. 19-73) provides additional guidance on the shapes of practice curves and conditions that cause those shapes to vary.

Checking for Stability

It is extremely important, before comparison of any stressor or experimental condition to baseline performance, to make sure that the baseline performance is "stable." Unless the point of stability has been reached, practice (which typically increases performance) will overlay stressor effects (which usually decrease performance) and the power of the study to detect effects that may be present can be sharply reduced.

There are three main criteria for stability of a test. First, the means should have begun to "level off" or approach asymptote. Second, the variances should be relatively stable from trial to trial. Third, and less well recognized, the correlations between trials should all be of about the same magnitude. Until correlational stability is achieved, individuals are still changing positions within the distributions of scores, that is, there are still subject by trial interactions, and overall stressor effects may be masked. The procedures for examining stability, particularly correlational stability, are complex, involve considerable exercise of judgment, and are sometimes tedious, but their outcomes provide a valuable tool for understanding the presence and absence of experimental or stressor effects. Appendix E gives a particularly thorough example of stability analysis.

Estimating Reliability

As noted repeatedly in previous sections, tests of low reliability provide only limited power for detecting stressor effects. It is important to estimate from the data the make sure that the baseline performance is "stable." Unless the point of stability has been reached, practice (which typically increases performance) will overlay stressor effects (which usually decrease performance) and the power of the study to detect effects that may be present can be sharply reduced.

There are three main criteria for stability of a test. First, the means should have begun to "level off" or approach asymptote. Second, the variances should be relatively stable from trial to trial. Third, and less well recognized, the correlations between trials should all be of about the same magnitude. Until correlational stability is achieved, individuals are still changing positions within the distributions of scores, that is, there are still subject by trial interactions, and overall stressor effects may be masked. The procedures for examining stability, particularly correlational stability, are complex, involve considerable exercise of judgment, and are sometimes tedious, but their outcomes provide a valuable tool for understanding the presence and absence of experimental or stressor effects. Appendix E gives a particularly thorough example of stability analysis.

Estimating Reliability

As noted repeatedly in previous sections, tests of low reliability provide only limited power for detecting stressor effects. It is important to estimate from the data the reliability of each test used in a study. Although Table 2 gives some "generic" reliability estimates for tests in the menu, reliabilities can be heavily impacted by the characteristics of a particular study (ability level and experience of subjects, test lengths selected, etc.). The reliabilities of interest are obtained from the intertrial correlations at and beyond stability points, and are estimated from the average of these correlations across all trials after stability and before introduction of the stressor condition. If there are large stressor or experimental effects, it can likely be inferred that tests are sufficiently reliable; if, however, stressor effects are absent or weak, it is important to know if such an outcome is due to problems with test reliability rather than to a real absence of effects.

8.0 SUMMARY

The first two chapters have led the user through the test menu, the selection of an appropriate battery for an application, and the configuring of the software to implement the battery. Recommended batteries and administrative guidance have been provided, the output files and scoring options have been described, and some screening analyses have been suggested. These two chapters comprise a free-standing "users manual" for a powerful test menu and flexible software. Much of the background information on which the specific recommendations are based has been deliberately omitted to keep the focus on the important procedures as clear as possible. The next chapter and its accompanying appendices describe the rationale of battery development and expand on the research and studies which support that development.

CHAPTER III -- RESEARCH UNDERLYING THE BATTERY

1.0 PAB BACKGROUND

The UTC-PAB is the product of the Tri-Service Joint Working Group on Drug Dependent Degradation of Military Performance (JWGD3 MILPERF), and is being designed as the primary instrument for Level II assessment of cognitive performance in a multiple-level drug evaluation program. The basic structure of the UTC-PAB evolved from a three-day, JWGD3 MILPERF-sponsored Task Area Group (TAG) workshop held in November 1984 at the Naval Medical Research Institute, Bethesda, Maryland, and was conceived by professionals with backgrounds in several content areas (e.g., sustained operations, information processing, workload assessment) and who were actively engaged in the development of performance batteries for specific applications in applied research. An indepth background of the Unified Tri-Service Cognitive Performance Assessment Battery (UTC-PAB) may be found in Englund, Reeves, Shingledecker, Thorne, Wilson, and Hegge (1987). Hardware and software design specifications have also been produced (Hegge, Reeves, Poole, & Thorne, 1985).

The thematic objective of the UTC-PAB development effort was to select tests from existing batteries and standardize on their design. This requirement for standarization included that they be written in common software. The proposed Performance Assessment Battery (PAB) includes 25 separate tests which emphasize information processing, cognition, memory, perception, and related mental acuity constructs. Recently, extensive documentation of the tests was compiled in a literature review (Perez, Masline, Ramsey, & Urban, 1987) which focused on the theoretical basis of each test, information regarding reliability, validity and sensitivity of the test, along with other specifications and subject instructions.

To our knowledge, the PAB tests have not previously been implemented on a portable computer or studied using a repeated-measures design for the purposes of evaluating stability, reliability, and correlations among tests. The only UTC-PAB study to date of which we are aware reports the results in two military pilot groups over 10 trials (Reeves & Thorne, 1988).

2.0 APTS BACKGROUND

The Automated Performance Test System (APTS) derives from a series of interlocking studies conducted by Essex Corporation, originally under NASA sponsorship and later including National Science Foundation support. From the outset this effort was keyed toward producing a battery-operated computer as a portable field unit for testing human performance after administration of ameliorative drugs for motion (space) sickness which might be potentially toxic. Philosophically, the current APTS effort built on an earlier program where repeated-measures analyses were conducted to create a menu of performance tests

(Performance Evaluation Tests for Environmental Research [PETER] - Kennedy & Bittner, 1977).

The philosophy of our approach to performance test development involves four different goals. The first is to deal with only tests or tasks that can be shown to be psychometrically sound. This requires that we demonstrate stability of means and standard deviations within few administrations, and most important, that correlational stability, the stability of trial-to-trial intercorrelations, be shown to occur quickly and with high test-retest prescreening correlations (i.e., reliability). The second goal is to demonstrate that the battery has factorial multidimensionality and that the subscales cross-correlate with earlier performance tests and other recognized instruments of ability. Third, it is necessary to demonstrate and document sensitivity to factors known to compromise performance potential in the laboratory and ultimately real-world situations. Fourth, the tasks must be shown to be predictive of the types of work performed in an operational context.

Environmental stressors are most often studied with a pre-, per-, post-paradigm. This approach makes maximum use of the "each subject serves as his or her own control" philosophy. As a practical matter, measures of operational performance are elusive and several problems remain in the assessment of human performance; chronically low retest reliability, instability across days due to learning, wide individual differences of unknown or uncontrolled variation, not knowing what to measure, etc. To obviate this problem test batteries of substitute tasks are often employed. Although it is difficult to get the world's "experts" to agree (Sanders, Haygood, Schroiff, & Wauschkuhn, 1986), it is our opinion that the two essential metric issues are "stability" and "reliability." The amount of time required can be critical in testing; therefore, tests which stabilize quickly and are reliable with less testing time are preferred over those which take longer.

The second requirement for meaningful and interpretable repeated measurements is that practice effects must be nil or predictable. Lord et al. (1968) point out that repeated measurements may be useful if mean scores change by an additive constant from one trial to another. Campbell and Stanley (1963), in their classic discussion, illustrate the principle that the additive constant should be the same across trials; the cumulative effect should have no more than a linear trend (preferably with near zero slope). They also noted that nonlinear changes across trials impede or make impossible interpretation of effects of experimental interventions.

The APTS Criteria

1. Stability -- Repeated measurements must possess certain characteristics to be meaningful and clearly interpretable

(American Psychological Association, 1974; Jones, 1972; Lord & Novick, 1968). First, the measurements must represent a constant mixture of human performance capabilities on each trial of repeated measurement. In its simplest form, this requirement implies that the relative differences between subjects, on the capability being measured, remain constant across all trials of repeated measurements. This requirement for meaningful repeated measurements can be met objectively by showing that, apart from measurement errors, intertrial correlations are unchanging (differentially stable) and variances are homogeneous across baseline repetitions (Bittner, 1979; Jones, 1980; Lord et al., 1968). Differential stability, in this context, provides assurance that the entity which is being measured is remaining constant (Alvares & Hulin, 1972). Stated technically, differential stability and constant variances make up the composed symmetry requirement of the variance-covariance for simple repeated-measures analysis of variance (Winer, 1971, p. 276-277). Together, differential and variance stability are required for simplified analysis and interpretation.

The requirement for differential stability distinguishes work conducted in the PETER and APTS programs from test battery development conducted by others. It is our view that unless tests have been practiced to the point of differential stability, attribution of effect (i.e., what the test tests) due to the experimental treatment is not possible.

In sum, the statistical requirements for easily interpretable results of repeated measures include level or linearly increasing means, level variances, and differential stability.

2. Stabilization Time -- Desirable performance measures should stabilize rapidly following brief periods of practice without forfeiting metric qualities. Any task under consideration for stressor or environmental research must be depicted in terms of the number of trials necessary to establish stability.

3. Task Definition -- Task definition is the average reliability of the stabilized task (Jones, 1979, 1980) and is calculated as the average intertrial correlation between testing trials following the trial when "differential stability" occurs. Higher average reliability (i.e., task definition) improves power in repeated-measures studies when variances are constant, because the lower the error within a measure the greater the likelihood that mean differences will be detected. Task definitions for different tests, however, cannot be directly compared without first standardizing tests for test length.

4. Reliability Efficiency -- Test reliability is known to be influenced by test length (Guilford, 1954); tests with longer

administration times and/or more items maintain a reliability advantage over shorter test times. Thus, test lengths must be equalized before meaningful comparisons can be made. A useful tool for making such relative judgments is called the reliability-efficiency, or standardized reliability, of the test (Kennedy, Carter, & Bittner, 1980), which is computed by correcting the reliabilities of different tests to a common test length or time by use of the Spearman-Brown prophecy formula (Guilford, 1954, p. 354). In our view, ability tests should not be considered to be reliable unless they reach $\underline{r} = 0.707$ for a three-minute session, which means that 50% of the variance is common across successive administrations.

5. Task Ceiling -- If all or several subjects obtain the maximum level of performance then the task is said to have a ceiling (Jones, 1980). Ceilings are undesirable because they limit discrimination between subjects and all those subjects perform equally well except for random error.

6. Factor Richness -- Finally, because different agents may interact with different aspects of performance, tasks which possess the features listed above should have minimum overlap; they should encompass as much unique variance as possible. Further, a battery of such tests should have as many factors as possible for a given testing time.

Performance Evaluation Tests for Environmental Research (PETER)

The PETER program was conducted at the Naval Biodynamics Laboratory in New Orleans, Louisiana, from 1977-1981. That work followed an "engineering" approach to test battery development -- it set out to evaluate the six metric properties (listed above) of tests BEFORE proposing them for inclusion and further consideration (Kennedy, Bittner, Harbeson, & Jones, 1981). In its early stages, virtually all the tests of that program were paper-and-pencil based or 35mm slide projector based. Later, video games (Jones, Kennedy, & Bittner, 1981) were employed.

The early framers of the PETER program took to heart criticisms about the drawbacks of following pschometrically derived theories of cognitive abilties (Carroll, 1974) and were therefore empirical in their approach. Except for the use of video games as tests, which was an innovation of that program, virtually all the other tests examined were drawn from existing batteries and/or the literature on experimental cognitive studies. The "ancestors" of the tests which served as subject matter for that work included Wechsler's Adult Intelligence Test (Wechsler, 1958); Halstead-Reitan Battery (Reitan & Davison, 1974); Episodic Memory Battery (Underwood, Boruch & Malmi, 1977); Information Processing Battery (Rose, 1974, 1978); Kit of Factor Referanced Cognitive Tests (Ekstrom, French, Harman, & Dermen, 1976; Moran, Kimble, & Mefferd, 1964); Manual Dexterity Battery (Fleishman & Ellison, 1962) and some miscellaneous tests

(Carter, Kennedy & Bittner, 1980). Although a selection battery was not the purpose of the PETER work, the "engineering" approach which was followed is consonant with advocacy of "process models instead of the traditional trait models" (Kyllonen, 1986).

The PETER program examined 114 tests (Bittner, Carter, Kennedy, Harbeson, & Krause, 1986) and determined their suitability for repeated-measures applications, 90 reports of that work are available (Harbeson, Bittner, Kennedy, Carter, & Krause, 1983). Approximately 30 tests were surfaced which met minimum suitability criteria for repeated-measures tests. The metric criteria which qualified a test for being suitable were: rapid stabilization (< 10 minutes' practice), high reliability (r > 0.707 for three minutes' testing), and no obvious ceiling.

Automated Performance Test Systems

In 1982, Essex obtained support from the National Aeronautical and Space Administration to mechanize a microcomputer-based battery of tests for use in the study of motion sickness preventatives. This work began with the 30 tests of the PETER program as the basis, since they had already demonstrated their requisite qualities for repeated measures tests. Later (1984), National Science Foundation support was obtained for a related effort -- development of a generic performance test battery for study of toxic chemicals and environments.

The two projects were coalesced into a series of interlcking experiments. These studies which are described below, have been published in a series of articles and technical reports (Appendix B), and included creation of software, computer implementation of tests, certification of tests in the new medium, and field trials of the portable units. In addition, several areas were which not addressed formally in the original PETER work (sensitivity, factor richness, and operational relevance) were to be studied experimentally. Also during this period, several laboratories purchased or borrowed systems and reports of these studies have been updated periodically through a series of newsletters.

General Hardware and Software Considerations

The tests of the PETER battery were initially implemented on a NEC PC8201A portable lap-top computer and became known as the Automated Performance Test System (APTS) (Bittner, Smith, Kennedy, Staley, & Harbeson, 1985). The 8201A was selected because of the amount of onboard memory available (64K bytes), and the low cost of the unit and peripherals (approximately $850.00 at the time of implementation). The display screen consisted of 240x64 pixel (40 characters by 8 lines) liquid crystal display (LCD) with adjustable contrast control. The unit is lightweight (3.8 pounds) and durable. Part of the work performed under the NASA contract was to present a system which would successfully clear minimum requirements for approval for flight on the Space Shuttle.

All tests for the original APTS are written in the BASIC software language. Many functions such as prompting for input, converting lower case letters to upper case, test timing, and response timing were common to all the tests. Assembly language programs were written to perform these common functions thereby providing more room in memory for data storage and the tests themselves.

Since the initial implementation of the test battery on the NEC, the IBM Personal Computer has become an industry standard, and the original test battery was converted for IBM-compatibles. Because the portability aspect of the test battery was a crucial feature, we selected the Zenith Data Systems ZFL-18X series as the current host of the portable assessment battery. The 18X contains 640K onboard memory, two 720K byte 3.5 inch floppy drives (or a 10 or 20 megabyte hard drive), serial and parallel interfaces, an RGB interface, and 80 characters by 25 line super twist, backlit LCD display, and is completely IBM PC compatible. The batteries are capable of powering the unit with both drives running and the brightness control set on high for 4.2 hours. From the present configuration, coversion to other portable systems (e.g., the new Paravant RHC-88) can be accomplished easily, although the price of the 18X series, its portability, and its ability to store large amounts of data make it an attractive device for field testing.

Psychometric Studies

For proof of concept, 20 subjects were tested in the first NASA sponsored study over four replications using paper-and-pencil versions as well as the computerized version of six tests (Kennedy, Wilkes, Lane, & Homick, 1985). All tests appeared to achieve stability within the four test sessions, reliability efficiencies were generally high (*r* >.707 for 3-minute testing), and the computerized tests were largely comparable to the paper-and-pencil version from which they were derived. The tests that were evaluated for inclusion in this experiment were Grammatical Reasoning, Pattern Comparison, Code Substitution, and the Tapping series, tests which had largely proven their metric properties in paper-and-pencil versions earlier in the PETER work. As these tests all exhibited stability and reliability within our proposed standards, all were proposed for further testing.

In the next NASA study (Kennedy, Wilkes, Dunlap, & Kuntz, 1987), in addition to evaluating stability and reliability of the tests, predictive validity was also examined. Twenty-five subjects were tested over significantly more replications (10) and tests (11) than previously. The 11 tests were concurrently administered in paper-and-pencil (marker battery) and microcomputer-based versions and compared to the Wechsler Adult Intelligence Scale (WAIS). Nine of the 11 microcomputer-based tests achieved stability. Reliabilities were generally high, with $r \geq .77$ for three minutes of testing for the recommended tests. Cross-correlations of microbased tests with traditional paper-and-pencil versions and indices of stability suggest equivalency between the tests in the different modes. Correlations between certain microbased subtests and the WAIS identified common variance.

In the third NASA study, also supported by NSF (Wilkes, Kuntz, Kennedy, & Tabler, 1988) 21 different tests, including six short-term memory tests which had not been studied before, were SELF-ADMINISTERED by the subjects without a full-time proctor. This experiment confirmed results from previous studies and demonstrated that self-administered tests are a viable alternative for repeated-measures study which may have application for toxic chemical and environmental testing. Air Combat Manuvering, Pattern Comparison, and Reaction Time Four-Choice took the longest of the original battery to stabilize. All tests stabilized by trial 5; the memory tests took a little longer and appeared to measure unique constructs but with only modest reliabilities.

A fourth study conducted a factor analysis and because it included some tests described in the UTC-PAB (Englund et al., 1987) it has been reproduced as Appendix C. In that work, 11 computerized tests from two performance test batteries were administered three times to each of 108 college students (48 males and 60 females). Factor analyses carried out on the total sample yielded three consistent factors: a cognitive complexity factor emphasizing encoding and analogical transformations on which Continuous Recall, Matrix Rotation, Grammatical Reasoning and Pattern Comparison load most heavily; a cognitive speed factor emphasizing the decoding of information and data entry on which Math Processing, Code Substitution and Pattern Comparison load most heavily; and a motor speed factor identified by the Tapping and Reaction Time tests. The Wonderlic Personnel Test was group administered before the first and after the last administration of the performance tests. The multiple R's in the total sample between combined Wonderlic as criterion and Grammatical Reasoning and Math Processing as predictors ranged between 0.41 and 0.52 on the three test administrations. Based on these results, a core battery was recommended consisting of two tests from each factor as time permits. This battery provides a reasonable, short estimate of IQ based on three well-identified factors; two information processing factors, one for input, throughput, and encoding and one for output and decoding, and the third a motor speed factor. This core battery can be usefully augmented, especially in operational situations, by other available tasks. Results are discussed in the context of the need to develop easily administered, psychometrically sound, factorially rich cognitive test batteries.

Sensitivity Studies

In addition to the development studies for purposes of test certification and microcomputer evaluation, a series of studies have been performed in collaboration with Essex personnel and by various laboratories using APTS tests. For example, NEC-based APTS batteries obtained by the USAF School of Aerospace Medicine were loaned to Navy scientists for use in the Persian Gulf, and to a USAF scientist who used them to study drugs and time zone effects. Some of these have been reviewed elsewhere (Kennedy, Lane, & Kuntz, 1987) and specific studies are referred to below:

Altitude -- Until recently, lack of an adequate human performance research tool has resulted in the employment of a variety of techniques, methods, and measures that limit systematic comparisons across altitude studies. Such limitations have delayed the development of a cohesive body of knowledge regarding human performance at altitude. Measurement and data collection inadequacies have further contributed to research difficulties. While highly controlled studies systematically relating sustained exposure to human performance are largely lacking, we believe that exceptions are beginning to appear (cf. e.g., Banderet & Burse 1984; Banderet, Benson, McDougall, Kennedy, & Smith, 1984).

The APTS battery has been tested at simulated altitude by scientists of the U.S. Air Force, and the U.S. Army Institute for Environmental Medicine (Banderet et al., 1984; Kennedy et al., 1988). The initial results show a definite cognitive performance decrement with sustained periods at altitudes of 23,000 feet, and with abrupt, short periods at 27,000 feet. However, motor performance remained essentially unchanged. An important point to note is that typical measures of performance would not have detected the effect altitude had on the mental capabilities of the participants.

Drugs -- With regular doses of certain motion sickness drugs, virtually all of the scores for both motor and cognitive tests changed in a theoretically rational direction in studies conducted by Dr. Charles Wood at Louisiana State University Medical School. That is, amphetamine scores increased and scopolamine scores decreased over placebo. A simple ANOVA revealed no significant outcomes (other than that Pattern Comparison, one of the APTS tests, scores appeared to be significantly poorer with hyoscine). The within-subject variables were scopolamine and dexedrine, arranged factorially in a totally within-subject design (a more powerful approach). The results indicate that amphetamine significantly increased Nonpreferred Hand Tapping (a motor skill test) and there was a trend for increased scores on the Sternberg (an item recognition test). This would mean there were more "hits" or that latency improved. There was not a significant effect of scopolamine on Preferred-Hand Tapping. The study further showed an interaction of scopolamine and dexedrine with Two-Hand tapping. Though not statistically significant, overall it appears that scopolamine facilitates performance more when dexedrine is also present then it does without dexedrine.

Chemoradiotherapy Treatments -- From the University of Washington, Dr. Parth has been studying patients who are receiving bone marrow transplants and chemoradiotherapy treatments. In this study, the tests of the basic NASA battery were administered, along with other tests, to both a patient population undergoing chemotherapy subsequent to bone marrow transplants and to a control population of sibling donors. Four replications of the battery were given spaced over one year, including prior to transplant therapy, during therapy, and in a follow-up examination. The primary purpose of such a study was to determine battery sensitivity to physiological stressors different from those examined in previous studies. The battery as a whole was strikingly effective in detecting performance shifts in patients and significantly differentiating patients from controls throughout the two therapy test periods. Greatest discrimination was apparent in the complex cognitive measures (i.e., Code Substitution) than in the "motor" (i.e., Tapping). Discrimination was present for both accessory and latency measures, although effects were stronger for accuracy performance. We have included a copy of a draft of this paper in Appendix D.

Sleep Loss -- Two different studies of sleep loss have been conducted. In the first study, Kiziltan (1985) at the U.S. Naval Postgraduate School in Monterey, California, observed statistically effects on Code Substitution but obtained only directional changes (nonsignificant) on the other tests following one night without sleep. Another study was performed with the NASA battery tests at Ames Research Center in Moffett Field, California. The experiment lasted 41 days, 30 of which was the bedrest phase. Tne results of this study revealed modest or no change on most tests.

In summary, for the past few years our research efforts have concerned study and identification of reliable performance measurement instruments for exotic environments. Under the sponsorship of NASA and NSF, a menu of performance tasks implemented on a battery-operated portable microcomputer has been developed. These measures differ from conventional performance measures in that tests need not involve operations in common with the performance measures, only components/factors in common. The tests also exhibit higher reliabilities ($r \geq .70$) than traditional field performance measures ($r = .10-.30$). Currently, 21 tests (some of which have 3 versions) are available on a menu for microcomputer presentation. These tasks are reliable and become stable in minimum amounts of time, appear sensitive to some agents, comprise constructs related to actual job tasks, and are easily administered and scored. Collectively these tests are known as the Automated Performance Test System (APTS). In numerous experiments the APTS has been shown to be a stable and reliable indicator of performance. If a person performs in a predictable manner and an intervening factor is introduced that has an adverse effect on performance (i.e., zero gravity, stress) is likely to be detected by one or more of APTS tests. Using a stable, sensitive, battery of performance tests would be analogous to taking a person's temperature, blood pressure, or weight. If administered on a daily basis it would be a form of record keeping that would show whether a person's performances were being affected by the environment or factors such as fatigue or workload. The APTS tests cognitive factors related to job performance and is therefore more predictive of performance than traditional methods of respiration, heart rate, blood pressure, et cetera.

3.0 DEVELOPMENT OF THE BATTERY

The chief reports of experiments conducted exclusively for the US Army Aeromedical Research Laboratory are included as appendices (E-G). The overall objective was to mechanize and implement the tests of the UTC PAB and then to evaluate them for repeated measures suitability according to the metric criteria listed above and to compare them to the marker tests of APTS.

appear sensitive to some agents, comprise constructs related to actual job tasks, and are easily administered and scored. Collectively these tests are known as the Automated Performance Test System (APTS). In numerous experiments the APTS has been shown to be a stable and reliable indicator of performance. If a person performs in a predictable manner and an intervening factor is introduced that has an adverse effect on performance (i.e., zero gravity, stress) is likely to be detected by one or more of APTS tests. Using a stable, sensitive, battery of performance tests would be analogous to taking a person's temperature, blood pressure, or weight. If administered on a daily basis it would be a form of record keeping that would show whether a person's performances were being affected by the environment or factors such as fatigue or workload. The APTS tests cognitive factors related to job performance and is therefore more predictive of performance than traditional methods of respiration, heart rate, blood pressure, et cetera.

3.0 DEVELOPMENT OF THE BATTERY

The chief reports of experiments conducted exclusively for the US Army Aeromedical Research Laboratory are included as appendices (E-G). The overall objective was to mechanize and implement the tests of the UTC PAB and then to evaluate them for repeated measures suitability according to the metric criteria listed above and to compare them to the marker tests of APTS.

Experimental Studies

a. In the first study (Appendix E) 25 right-handed males from the University of Central Florida were tested with six tests from the Unified Tri-Service Cognitive Performance Assessment Battery and six tests from the Automated Performance Test System (APTS). Task descriptions of all the tasks are in Appendix A. From the standpoint of repeated-measures applications, most tests stabilized with very nearly acceptable levels of retest reliability. However, the APTS tests usually fared better than candidate PAB tests despite differences in test administration. Also, PAB tests are accompanied by feedback and were preceded by a one-hour orientation period in which subjects became familiar with the microcomputer and could ask questions regarding task instructions. However, APTS tests were not introduced until the 8th day of testing, which could suggest that a generalized learning curve may have asymptoted by the time the new tests were introduced. It was considered that the recommended PAB training times (Englund et al., 1987) may be underestimated. These issues, as well as the results of the individual tests appear in Appendix E.

b. The second study under this contract followed essentially the same paradigm as the first and employed twenty-five males from the University of West Florida. Five tests were selected from the PAB and six performance tests from the APTS. PAB tests were also selected because they were comparable to tests within the APTS battery. For example, both batteries had Grammatical Reasoning, Code Substitution, and Four-Choice Reaction Time, though they were of different versions. Both batteries contained spatial rotation tests with the APTS' Manikin and PAB's Matrix Rotation. These tests are similar in that they involve different aspects of spatial transformation. However, if tests with common names are demonstrated to be sufficiently isomorphic, then the large literature behind the different versions can be pooled. Comparing similar tests across studies assists in defining a universal test taxonomy. In addition, a series of 10-second finger tapping exercises was included in both test batteries as a check against intervening factors during battery administration.

Most tests stabilized with very nearly acceptable levels of retest reliabilities; the APTS tests were generally "better" than PAB on reliability and stability when adjusted for test length. The problem encountered by Turnage et al. (1987), where the participants were not exposed to APTS until the eighth day of testing, was avoided in this study as all participants took both batteries daily from the first session. The specific findings from the two sets of tests may be found more completely described in Appendix F.

c. The third study (Appendix G) represents an extension of the Turnage et al. (1987) and Tabler et al. (1987) studies of the psychometric properties of PAB and APTS to evaluate an additional five PAB tests which were considered suitable for microcomputer administration. In addition, the study utilized a "smart" repeated-measures system to detect and warn subjects who were responding below criteria for accuracy. Twenty-five students, (nine males and 16 females) from the University of Central Florida were recruited for participation in this study. Eight performance tests were selected from the PAB and five from the APTS. Three of the PAB tests had been evaluated in early studies in this series. Feedback (knowledge of results) as to performance was furnished to participants during the orientation session but not during the ensuing sessions.

The psychometric comparability between PAB and APTS tests in this study, in which both number correct and response latency response measures achieved high levels of acceptability, reflects improvements in microcomputer test mechanization or implementation, as well as procedural techniques. PAB test properties were materially improved by the new procedures. In the Turnage et al. study (1987), APTS tests were not introduced until the eighth day of testing, and the overall estimated reliabilities for PAB and APTS batteries were 0.60 and 0.88, respectively, for number correct and response latency measures. These figures suggest that a generalized learning curve may have asymptoted by the time the APTS was introduced. In the Tabler et al. (1987) study, all participants took both batteries daily from the first session but the average estimated reliabilities for PAB and APTS batteries, using number correct and response latency measures, were still divergent at 0.64 and 0.84, respectively. It is our view that the comparability of the two test batteries evidenced in the current study may well be a function of the implementation of the "smart" subroutine during training which flagged the experimenter when it appeared that instructions were not clearly understood. If PAB tests require more training time prior to testing than APTS tests, then it could be assumed that participants entered repeated-testing sessions with comparable expertise on each battery, thus equalizing intertrial correlations. This hypothesis will be tested in a future planned study in this series, wherein the exact number of retrials prompted by the "smart" system during training will be automatically recorded.

In this third study the average correlation between number correct and response latency measures was -1.00 for APTS tests (corrected for attenuation) and -0.83 for PAB tests. The magnitude of this correlation indicates that, taking the size of the relationship into account, the two measures are redundant: persons with the shortest latencies have the most hits and the converse. This observation suggests that in future studies, provided that tests are of a fixed time, either response measure may be used to make generalizations about the other without seriously compromising research conclusions and at great savings in data reduction, analysis, and inferences. However, we advocate recording all these measures (hits, latency, and percent correct) for post hoc analysis.

In the three repeated-measures analyses of selected tests from PAB and APTS reported in this and previous studies, we have used a relatively elementary method to determine the factorial purity of individual tests (i.e., apportioning intertest correlations into three levels of overlap). It is our observation that some tests interrelate more than others by virtue of their intended measurement of similar psychomotor and cognitive constructs as well as their demonstrated psychometric similarity. Thus, additional testing was directed toward understanding the factor structure of tha PAB and APTS batteries.

d. In a fourth study aimed primarily at obtaining a larger sample size on more variables to more clearly observe the dimensionality of the combined battery, two response measures (number correct and response latency) were collected across five testing sessions on 26 tests for 100 students at the University of Central Florida. Tests included the complete battery of testable PAB tests as well as the most commonly used APTS marker tests. Testing times were abridged from those in earlier trials, both to enable a manageable session length and to explore the reliabilities of the tests in relatively short testing sessions (one to two minutes in most cases). Either number correct or response latency was used for each test, but not both. Preliminary factor analyses were performed. The results of those analyses were combined with those of earlier analyses and are presented within the factor structure in Table 3. Outcomes largely confirm the findings of the four NASA/NSF studies and the first three studies of the present development. The battery appears to tap processes or functions which lend themselves to interpretation along dimensions of information processing abilities rather than toward more traditional "trait" models. These dimensions are dynamic, and their importance in test performance is modifiable with practice. More detailed analyses in progress will address questions such as shortest reliable test length, more rigorous examination of factor structure changes across repeated-testing trials, the presence of general and subgroup factors in the structure, and whether the factor structure is similar for both number correct and response latency measures. Outcomes of these analyses will be reported separately.

Reliabilities obtained in the fourth study were somewhat lower than those in previous analyses, due primarily to the shorter test lengths. Findings lead us to recommend that the PAB tests should generally be run for three minutes (at least 2.5), and the APTS tests for at least two minutes (not including the tapping and reaction time series).

REFERENCES

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

American Psychological Association (1974). Standards for educational and psychological tests. Washington, DC: American Psychological Association.

Bittner, A. C., Jr. (1979). Statistical tests for differential stability. Proceedings of the 23rd Annual Meeting of the Human Factors Society (pp. 541-545). Santa Monica, CA: Human Factors Society. Also, Naval Biodynamics Laboratory, New Orleans, LA, September 1981, pp. 10-14 (Research Rep. No. NBDL-81R010) (NTIS No. AD A111086)

Bittner, A. C., Jr., Carter, R. C., Kennedy, R. S., Harbeson, M. M., & Krause, M. (1986). Performance Evaluation Tests for Environmental Research (PETER): Evaluation of 114 measures. Perceptual and Motor Skills, 63, 683-708.

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

Carroll, J. B. (1974, May). Psychometric tests as cognitive tasks. A new "structure of intellect" (Tech. Rep. No. 4, ETS-RB-74-16). Washington, DC: Office of Naval Research, Personnel and Training Research Programs.

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. Also, New Orleans, LA: Naval Biodynamics Laboratory, July 1981, pp. 1-7 (Research Rep. No. NBDL-81-R0088). (NTIS No. AD A111296)

Englund, C. E., Reeves, D. L., Shingledecker, C. A., Thorne, D. R., Wilson, K. P., & Hegge, F. W. (1987). Unified Tri-Service Cognitive Performance Assessment Battery (UTC-PAB): I. Design and specification of the battery (Rep. No. 87-10). San Diego, CA: Naval Health Research Center.

Guilford, J. P. (1954). Psychometric methods (2nd ed.). New York: McGraw-Hill, 400-402.

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

Jones, M. B. (1972). Individual differences. In R. N. Singer (Ed.), *The psychomotor domain* (pp. 107-132). Philadelphia, PA: Lee & Febiger.

Jones, M. B. (1979). *Stabilization and task definition in a performance test battery* (Final Rep., Contract N00203-79-N-5089). New Orleans, LA: U.S. Naval Aerospace Medical Research Laboratory.

Jones, M. B. (1980). Sequential precision and diminishing returns in the acquisition of a motor skill. *Journal of Motor Behavior*, *12*, 69-73.

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)*. Personnel Performance Measurement Symposium, San Diego, CA. L. T. Pope & D. Meister (Eds.) *Productivity enhancement: Personnel performance assessment in Navy systems*. Navy Personnel Research & Development Center. (NTIS No. AD A056074)

Kennedy, R. S., Wilkes, R. L., Dunlap, W. P., & Kuntz, L. A. (1987, October). *Microbased repeated-measures performance testing and general intelligence*. Paper presented at the 29th Annual Conference of the Military Testing Association, Ottawa, Ontario, Canada.

Kennedy, R. S., Wilkes, R. L., Lane, N. E., & Homick, J. L. (1985). *Preliminary evaluation of microbased repeated -measures testing system* (Report No. EOTR-85-1). Orlando, FL: Essex Corporation.

Kyllonen, P. C. (1986). *Theory-based cognitive assessment* (Rep. No. AFHRL-TP-85-30). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Lord, M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Redding, MA: Addison-Wesley.

Perez, W. A., Masline, P. J., Ramsey, E. G., Urban, K. E. (1987). *Unified Tri-Service Cognitive Performance Assessment Battery: Review and methodology* (Rep. No. AAMRL-TR-87-C07). Wright-Patterson AFB, OH: Armstrong Aerospace Medical Research Laboratory.

Reeves, D. L., & Thorne, D. P. (1988). Development and application of the Unified Tri-Service Cognitive Assessment Battery within naval aviation. Paper presented at the 59th Annual Scientific Meeting of the Aerospace Medical Association, New Orleans.

Rose, A. M. (1978). An information processing approach to performance assessment (Rep. No. AIR 58500-11/78-FR). Washington, DC: American Institutes for Research.

Tabler, R. E., Turnage, J. J., & Kennedy, R. S. (1987). Repeated-measures analyses of selected psychomotor tests from PAB and APTS: Stability, reliability, and cross-task correlations: Study 2. Orlando, FL: Essex Corporation.

Turnage, J. J., Kennedy, R. S., & Osteen, M. K. (1987). Repeated-measures analyses of selected psychomotor tests from PAB and APTS: Stability, reliability, and cross-task correlations. Orlando, FL: Essex Corporation.

Underwood, B. J., Boruch, R. F., & Malmi, R. A. (1977, May). The composition of episodic memory (ONR Contract No. N00014-76-C-0270). Evanston, IL: Northwestern University. (NTIS No. AD A040696).

Wechsler, D. (1958). Measurement and appraisal of adult intelligence (4th ed). Baltimore, MD: Williams & Wilkins.

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

APPENDIX A

APTS and PAB TEST DESCRIPTIONS

Automated Performance Test System

Associative Memory
Code Substitution
Complex Counting
Grammatical Reasoning
Manikin
Mood Adjective Checklist
Number Comparison
Pattern Comparison
Reaction Time
Sternberg
Tapping

Performance Assessment Battery

Code Substitution
Continuous Recall
Grammatical Reasoning
Grammatical Reasoning (Symbolic)
Item Order
Linguistic Processing
Manikin
Mathematical Processing
Memory Search
Matrix Rotation
Neisser
Pattern Comparison (Simultaneous)
Pattern Comparison (Successive)
Reaction Time
Spatial Processing
Stroop
Time Wall
Vertical Addition
Visual Vigilance

AUTOMATED PERFORMANCE TEST SYSTEM

Associative Memory

This is a memory test (Underwood, Boruch, & Malmi, 1977) that requires the participant to view five sets of three letters that are numbered 1 to 5 and memorize this list. After an interval, successive trigrams are displayed and the participant is required to press the key of the number corresponding to that letter set.

Code Substitution

Code Substitution is described as a cognitive and perceptual-type task with visual search encoding and decoding, rote recall, and perceptual speed as important factors in performance. The computer displays nine characters across the top of the screen, and beneath them the digits 1 through 9 within parentheses. The participant is to associate the digit with the character above it. This is called the participant's "code." Under the code are two rows of characters with empty parentheses beneath them. The participant is required to insert the number associated with the character from the code above via the corresponding key press. When the participant has completed a row, a new row scrolls up to fill the position. According to Bittner et al. (1986), "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" (p. 38).

Complex Counting

Visual Counting: 1, 2, or 3
Auditory Counting: 1, 2, or 3

The Counting tests (Jerison, 1955) require the subject to accurately monitor the repeated occurrence of a particular stimulus. These tests require vigilance skills incorporated with a workload factor. The participant is required to count the number of times a box (visual) or tone (auditory) occurs. There are three different cues, boxes for the visual, referred to as left, middle, and right, and three tones for the auditory, identified as low, medium, and high. In the low demand task, the participant is to respond to every fourth low tone/left box and then press the left arrow key. The medium demand version of the task requires the subject to count not only the low tones/left boxes, but also the middle tones/boxes, and press the middle arrow key after every fourth middle cue. In the high demand version of the test, the participant must count each low, each middle, and each high cue and press the corresponding arrow key for every fourth low, every fourth middle, and every fourth high cue.

Grammatical Reasoning

The Grammatical Reasoning test (Baddeley, 1963) employs 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 verbs "follows" and "trails," and (5) A versus B mentioned first. There are 32 possible items arranged in random order. The participant's task is to respond "True" or "False," depending on the verity of each statement, by hitting the "T" or "F" keys respectively. Grammatical Reasoning is described as the measuring of higher mental processes with reasoning, logic, and verbal ability being the important factors in test performance (Carter, Kennedy, & Bittner, 1981).

Manikin

The Manikin Test (Benson & Gedye, 1963) involves the presentation of a simulated human figure (a sailor) in either a full-front or full-back facing position. The figure is shown holding three hearts, diamonds, clubs, or spades. One of the two patterns held matches a pattern which appears below the figure and is contained within the podium on which the figure is standing. The participant must indicate which hand is holding the matching pattern shown on the podium by pressing the appropriate arrow key. Pattern type, hand associated with the matching pattern, and front-to-back figure orientation (i.e., the figure is either facing or has its back to the participant), are randomly determined for each trial. The Manikin Test is a perceptual measure of spatial transformation of mental images and involves spatial ability (Carter & Woldstad, 1985).

Mood Adjective Checklist

This questionnaire gives an indication of the participant's mood at the time of administration. The participant chooses between "does not apply," "somewhat applies," and "definitely applies" to 15 out of 50 randomly generated adjectives.

Number Comparison

The Number Comparison task (Ekstrom, French, Harman, & Dermen, 1976) involves the presentation and comparison of two sets of numbers. The participant's task is to compare the numbers and decide if they are the same or different. Number sets may range from three to seven digits in length with the second number set always equal in digits to the first, and only one digit in the second set may be different from the first set of numbers. Number comparison has been described as a perceptual task that measures perceptual speed.

Pattern Comparison Simultaneous

The Pattern Comparison Test (Klein & Armitage, 1979), which measures factors relating to target acquisition and visual search, requires the participant to examine a pair of eight-dot patterns and to determine whether they are the "same" or "different." Patterns are randomly generated with similar and different pairs presented in random order. Performance is scored according tc the number of pairs correctly identified as similar or different. Pattern Comparison has been described as a spatial ability important to perceptual performance.

Reaction Time

1 Choice
2 Choice
4 Choice

The Visual Reaction Time Test (Donders, 1969) involves the presentation of a visual stimulus and measurement of a response latency to the stimulus. The participant's task is to respond as quickly as possible with a key press to a simple visual stimulus. On this test 1, 2, or 4 (depending on the number of choices) "outlined" box(es) are displayed above the 1, 2, 3, and 4 number keys. A short tone precedes at a random interval to signal a "change" in the status of the box(es) is about to occur. The box changes from "outlined" to "filled." The participant cbserves the box(es) for the change and ther presses the function key beneath the box that does change. Simple reaction time has been described as a perceptual task responsive to environmental effects (Krause & Bittner, 1982).

Sternberg (Short-Term Memory)

The Short-Term Memory Task (Sternberg, 1966) involves the presentation of a set of four letters for one second (positive set), followed by a series of single letters presented for two seconds (probe letters). The participant's task is to determine if the probe letters accurately represent the positive set and respond with the appropriate key press. Subject response is recorded from the two buttons (T-true, F-false) on the keyboard. Performance is based on the number of probes correctly identified. Short-Term Memory is described as a cognitive task which reflects short-term memory scanning rate (Bittner et al., 1986).

Tapping

Preferred hand tapping
Nonpreferred hand tapping
Two finger tapping

The tapping tests are motor skill performance tasks that may be placed at the beginning and at the end of a test battery, serving as a check against interfering factors during battery administration (i.e., boredom). The participant is required to press the indicated keys as fast as he or she can with either the preferred or nonpreferred hand or with the index fingers from both hands. Performance is based on the number of alternate key presses made in the allotted time. In a recent study (Kennedy, Wilkes, Lane, & Homick, 1985), tapping was described as a psychomotor skill assessing factors common to both Aim and Spoke.

PERFORMANCE ASSESSMENT BATTERY

Code Substitution Test

Adapted from a paper-and-pencil version of the test contained in the Wechsler Adult Intelligence Scale from Wechsler (1958), this test is designed to measure associative learning ability and perceptual speed. A string of nine letters and nine digits (numbers) are displayed across the screen in an arrangement so that the digit string is immediately below the letter string. Letters and digits are randomly paired for each test and their order is randomly assigned in the coding string. A test letter is presented at the bottom of the screen below the coding strings. The participant is to indicate which digit corresponds to that test letter in the display strings. The letter and digit associates change at 10-second intervals.

Continuous Recall

Continuous Recall (Hunter, 1975) measures the ability to encode and store information in working memory. The test consists of a random series of visually presented numbers which must be encoded by the participant in a sequential fashion. As each number is presented a "probe" number is simultaneously presented. The participant must compare this "probe" to a previously presented number at a prespecified position back in the series. Once the recall has been made, the participant must decide if that number is the same (S) as or different (D) from the "probe." The test can be made more difficult by using numbers comprised of several digits.

Grammatical Reasoning

Adapted from Baddeley (1968), this test is designed to measure logical reasoning ability (the integration and manipulation of information). Stimulus items are sentences of varying syntactic structure (i.e., A precedes B) accompanied by a set of letters (i.e., A, B). The single sentence problems are comprised of all possible combinations (32) of five binary conditions: (1) active versus passive wording, (2) positive versus negative wording, (3) key word "follows" versus "precedes," (4) order of appearance of the two symbols within the sentence, and (5) order of the letters in the simultaneously presented symbol set. The sentence must be analyzed by the participant to determine whether it correctly describes the sequence of the symbols in the symbolic set which appears to the right of the sentence. The lowest level of the test demand conditions was selected, which provided only a single sentence (e.g., A follows B). If the order is correctly described (true), the participant should press the "T" key. If the order is incorrectly described (false), the "F" key should be pressed. The low demand condition is 2.5 seconds.

Grammatical Reasoning (Symbolic)

The symbolic version of Grammatical Reasoning has three levels of difficulty, and uses symbols (i.e., *, @, and #) instead of the letters "A" and "B". Sentences of different syntactic structure accompanied by a set of symbols are presented simultaneously. Participants respond by selecting either "T" for true and "F" for false. The low demand task condition which was used in this study presented single-sentence items of variable syntactic construction that describe the order of pairs of symbols (all possible stimuli in the traditional version). The stimulus population for single-sentence problems is comprised of all possible combinations (32) of the following five binary conditions: (1) active versus passive wording of sentences; (2) positive versus negative wording; (3) keyword "follows" versus "precedes"; (4) order of the two symbols in the sentence; and (5) order of symbols in the symbol set. For the low demand condition the response deadline is 2.5 seconds.

Item Order

Item Order (Wilson, Pollack, & Wallick, 1986) is a test of short-term memory. A set of seven consonants are displayed on the screen for two seconds. After a predetermined pause, a new set of letters is presented. The participant must indicate if this second set of letters is identical to the first. Both sets must have all the same letters as well as having the letters in the same position to be considered identical. The response keys are "S" (same) and "D" (different).

Linguistic Processing

Physical Letter Match (low demand)
Category Match (moderate demand)

This task requires the participant to process linguistic information and classify letter pairs (Craik & Tulving, 1975; Posner & Mitchell, 1967). The participant must determine if letter pairs presented simultaneously match on a specified dimension. Two levels of difficulty are used: physical letter match (low demand) and category match (moderate demand). Physical letter match requires that the letter pairs presented are physically identical in order to constitute a match and has a response latency of 1.0 seconds. Category match specifies that a match occurs only when both letters of the pairs are either vowels or consonants and has a response latency of 1.5 seconds.

Manikin

Originally developed by Benson and Gedye (1963), this test is designed to index the ability to mentally manipulate objects and determine orientation of a given stimulus. The manikin is a sailor standing on a pedestal on which three hearts, diamonds, clubs, or spades appear (the matching stimulus). The figure is shown holding a box in each hand. Inside the boxes will appear three hearts, diamonds, or clubs (the comparison stimuli). The objective of this task is to determine which hand (right or left) matches the objects that appear on the pedestal on which the sailor is standing. The participant indicates an answer by pressing one of two arrow keys. The manikin may appear either upright or upside down and facing either toward or away from the participant. The manikin is centered in the middle of the display area and occupies approximately the full height available. The manikin is clothed in such a way as to make the front and back easily discriminable, using details such as facial features, and collar back. The figure remains displayed until the participant makes a valid right or left response by pushing the corresponding arrow key on the keypad or the response deadline is reached.

Mathematical Processing

Low Demand - Single Operator (+,-), 1.5 sec
Moderate Demand - Two Operator (+ -, - +, - -), 3 sec
High Demand - Three Operator (+ + -, + - -, - + -), 4 sec

Mathematical Prccessing (Shingledecker, 1984) requires the participant to perform arithmetical operations as well as value comparison of numeric stimuli. The participant performs one to three addition and/or subtraction operation(s) in a single presentation. A response is then made which indicates whether the total is greater or less than a prespecified value of five using the arrow keys. The problems are randomly generated using only numbers 1 through 9. The response deadline varies corresponding to the demand characteristic of the test as noted above.

Matrix Rotation

This test (Phillips, 1974) assesses spatial orientation and short-term memory. A series of 5x5 cell matrices are presented (singly) that contain five illuminated cells per matrix. The participant compares successive displays to determine if they are the same ("S") or different ("D"). Matrices are considered alike if the same matrix is rotated either 90 degrees to the left or 90 degrees to the right from the previously displayed matrix. Two successive matrices are never presented in exactly the same orientation.

Memory Search (Modified Sternberg)

Visual-Fixed Set
Visual-Mixed Set
Visual-Varied Set

This task (Sternberg, 1969) requires a participant tc maintain in memory a "study set" of alphabetic characters and to indicate whether the probe letters presented are in the study set or not. A trial consists of the presentation of one study set containing four letters followed by 10 probe letters. Although precise training items for this task has not been determined, Englund et al. (1986) suggest that major practice effects are eliminated within seven to sixteen trials based upon extrapolation from similar tests.

Neisser (Visual Scanning)

This test, adapted from Neisser (1963), is designed to measure visual search and recognition. The participant must scan an area or array of distractor objects in search of a target object. In this task the target and distractor objects are letters of the 26-letter alphabet arranged as 25 rows and five columns. The distractors include the 26 letters A through Z, excluding the letter K. A randomly selected character within the array is then replaced by the target letter K, with the restriction that the target letter may not occur within the first three or the last row. The array does not scroll or sweep but appears within a one-frame interval. The participant is instructed to scan the array in a normal reading sequence (left to right, top to bottom) and press any key upon detection of the target letter K. Once a response is made, the participant has 10 seconds to type in the two-digit row number which contains the stimulus character.

Pattern Comparison (Successive)

This test is designed to measure visual pattern recognition and spatial memory (Thorne et al., 1985). A random pattern of "dots" is displayed briefly on the screen and then followed after a blank retention interval by a second pattern that may be the same or different. The "dots" are depicted as white asterisks and the pattern is displayed for 1.5 sec. The screen blanks for three seconds and the second pattern is displayed until the subject responds or a 15-second deadline is reached. The participant decides whether the second pattern is the same or different as quickly as possible and presses the "S" or "D" key.

Pattern Comparison (Simultaneous)

The Pattern Comparison Test (Klein & Armitage, 1979) is designed to measure perceptual speed, an aspect of spatial ability. The participant views two eight-dot patterns that are displayed adjacent to each other. The task is to determine whether or not the two patterns match and respond by pressing one of two buttons, "S" for same or "D" for different. Two eight-dot patterns are enclosed inside borders that form a box around each pattern.

Reaction Time

This test is a derivative of a task developed by Wilkinson and Houghton (1975) and assesses the participant's ability to encode and categorize information, as well as their ability to select a response and execute a reaction. The test consists of the presentation of a flashing plus sign (+) imposed on a cursor in one of four quadrants of the CRT. The task involves pressing "the arrow key" (one of four directions) on the keyboard which corresponds to the quadrant containing the flashing plus sign. The stimulus remains in a quadrant until a response key is pressed and then reappears randomly within one of the quadrants. If one of the four keys is not pressed within a 2.5-second time period, the computer beeps at 0.1-second intervals until a response is elicited.

Spatial Processing

Two Bar, 0 degree rotation
Four Bar, 90 degree rotation
Six Bar, 180 degree rotation

This task (Chiles, Alluisi, & Adams, 1968) requires the participant to determine if the first histogram presented is the same or different from the following histogram. The second histogram in the pair may be rotated depending on the condition of the demand. Levels of difficulty include: two-bar histogram with a 0 degree angle; four-bar histogram that may be rotated 90 degrees, and six-bar histograms that may be rotated 180 degrees.

Stroop

Control Condition
Interference Condition
Combined Condition

This task was derived from Stroop (1935) and measures susceptibility to response competition interference. Three different versions may be used. All versions use the words "red," "blue," and "green," as well as the colors red, blue, and green. In the Control Condition, the word and the color displayed are congruent; the participant presses the key that represents that color. In the Interference Condition, the words and colors are usually, but not always incongruous; the participant then presses the key that represents the display color. In the Combined Condition, a neutral word (house, gun, door) is displayed in a particular color and the subject must press the key that represents that color.

Time Wall

This is a nonverbal time estimation task (Seppala & Visakorpi, 1983) in which a small object descending at a constant velocity passes behind a barrier. The task is to estimate when the object will reach the bottom edge of the barrier. The barrier contains a box which is the same shape and size as the object and the participant estimates the moment when the entire notch will be filled. This implementation uses a nominal 10-second time interval. The barrier occupies the lower third of the display area. The notch is centered along the wall's bottom edge. The moving object emerges from the top of the display area and descends at a constant velocity such that its leading edge would reach the bottom line of the display at a precisely known time (10 seconds). The falling box appears to pass behind the barrier, after which the timer continues to run but nothing else occurs until the participant responds or a deadline elapses. The participant estimates the transit time of the falling box and presses a designated response key. Feedback that an acceptable response has been made is provided by instantly filling the notch with the wall color.

Vertical Addition

This test is a two-column addition task (Ekstrom, French, Harman, & Dermen, 1976) that measures the ability to sum simple addition problems with speed and accuracy. In this test, a set of three two-digit numbers are simultaneously presented in a column format. The participant is required to sum as rapidly as possible and enter the answer via the keypad. The column of digits disappear with the first valid key entry and the trial ends when the return key is pressed or when a period of 30 seconds elapses.

Visual Vigilance

This is a vigilance test that corresponds to Donders' (1969) reaction time. The test simulates skills required of radar operators, word processors, and air traffic controllers. The participant searches for either the letter "A" or the number "3" in a random series of letters and numbers that are individually flashed on screen at random intervals. As soon as an A or 3 is identified the participant is to press any key on the keyboard. Though this test can be made any length, the longer the test is made, the more closely it simulates actual vigilance.

REFERENCES

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

Benson, A. J. & Gedye, J. L. (1963). Logical processes in the resolution of orientation conflict (Report 259). Farnborough, UK: Royal Air Force, Institute of Aviation Medicine.

Bittner, A. C., Jr., Carter, R. C., Kennedy, R. S., Harbeson, M. M., & Krause, M. (1986). Performance Evaluation Tests for Environmental Research (PETER): Evaluation of 114 measures. Perceptual and Motor Skills, 63, 683-708.

Bittner, A. C., Jr., Carter, R. C., Krause, M., Kennedy, R. S., & Harbeson, M. M. (1983). Performance tests for repeated measures: Moran and computer batteries. Aviation, Space, & Environmental Medicine, 54, 923-928.

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., Kennedy, R. S., Bittner, A. C., Jr., & Krause, M. (1980). Item recognition as a performance evaluation test for environmental research. Proceedings of the 24th Annual Meeting of the Human Factors Society (pp. 340-344). Santa Monica, CA: Human Factors Society.

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

Carter, R. C., & Wolstad, J. C. (1985). Repeated measurements of spatial ability with the Manikin test. Human Factors, 27(2), 209-219.

Chiles, W. D., Alluisi, E. A., & Adams, O. S. (1968). Work schedules and performance during confinement. Human Factors, 10(2), 143-169.

Donders, F. C. (1969). On the speed of mental processes. (Translated by W. G. Koster) Acta Psychologica, 30, 412-431.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976, August). Manual for kit of factor-referenced cognitive tests (Office of Naval Research Contract No. N00014-71-C-0117). Princeton, NJ: Educational Testing Service.

Englund, C. E., Reeves, D. L., Shingledecker, C. A., Thorne, D. R., Wilson, K. P., & Hegge, F. W. (1986). Unified Tri-service Cognitive Performance Assessment Battery (UTC-PAB) (Rep. No. 86-1). Fort Detrick· MD: U.S. Army Research and Development Command.

Hunter, D. R. (1975). Development of an enlisted psychomotor/perceptual test battery (AFHRL-TR-75-60). Wright Patterson Air Force Base, OH: Air Force Human Resources Laboratory.

Jerison, H. J. (1955, December). Effect of a combination of noise and fatigue on a complex counting task (WADC TR-55-360). Wright-Patterson Air Force Base, OH: Wright Air Development Center, Air Research and Development Command, United States Air Force.

Kennedy, R. S., Dunlap, W. P., Jones, M. B., Lane, N. E., & Wilkes, R. L. (1985). Portable human assessment battery: Stability, reliability, factor structure, and correlation with tests of intelligence (Tech. Rep. No. EOTR-86-4). Orlando, FL: Essex Corporation.

Kennedy, R. S., Wilkes, R. L., Lane, N. E., & Homick, J. L. (1985). Preliminary evaluation of a microbased repeated-measures testing system (Tech. Rep. No. EOTR-85-1). Orlando, FL: Essex Corporation.

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

Krause, M., & Bittner, A. C., Jr. (1982). Repeated measures on a choice reaction time task (Res. Rep. No. NBDL-82R006). New Orleans: Naval Biodynamics Laboratory. (NTIS No. AD A121904)

Krause, M., & Kennedy, R. S. (1980). Performance Evaluation Tests for Environmental Research (PETER): Interference susceptibility test. Proceedings of the 7th Psychology in the DoD Symposium (pp. 459-464). Colorado Springs, CO: USAF Academy.

Neisser, U. (1963). Visual search. Scientific American, 210(6), 94-103.

Pepper, R. L., Kennedy, R. S., Bittner, A. C., Wiker, S. F., & Harbeson, M. M. (1965). Performance Evaluation Tests for Environmental Research (PETER): Code substitution test. Perceptual and Motor Skills, 61, 735-745.

Phillips, W. A. (1974). On the distinction between sensory storage and short-term visual memory. Perception and Psychophysics, 6, 283-290.

Seppala, T., & Visakorpi, R. (1983). Psychophysiological measurements after oral atropine in man. Acta Pharmacologica and Toxicologica, 52(1), 68-74.

Shingledecker, C. A. (1984). A task battery for applied human performance assessment research (Tech. Rep. No. AFAMRL-TR-84). Dayton, OH: Air Force Aerospace Medical Research Laboratory.

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

Stroop, J. R. (1935). Factors affecting speed in serial verbal reactions. Psychological Monographs, 50, 38-48.

Thorne, D. R., Genser, S. G., Sing, H. C., & Hegge, F. W. (1985). The Walter Reed Performance Assessment Battery. Neurobehavioral Toxicology & Teratology, 7, 415-418.

Underwood, R. S., Boruch, R. F., & Malmi, R. A. (1977, May). The composition of episodic memory (ONR Contract No. N00014-76-C-0270). Evanston, IL: Northwestern University. (NTIS No. AD A040696)

Wechsler, D. (1958). Measurement and appraisal of adult intelligence (4th ed.). Baltimore: Williams and Wilkins Company.

Wilkes, R. L., Kennedy, R. S., Dunlap, W. P., & Lane, N. E. (1986). Stability, reliability, and cross-mode correlation of tests in a recommended 8-minute performance assessment battery (Tech. Rep. No. EOTR-86-4). Orlando, FL: Essex Corporation.

Wilkinson, R. T., & Houghton, D. (1975). Portable four-choice reaction time test with magnetic tape memory. Behavior Research Methods & Instrumentation, 7(5), 441-446.

Wilson, K. P., Pollack, J. G., & Wallick, M. T. (1986). The effects of ship motion on human performance. Proceedings of the American Society of Naval Engineers (pp. 347--395). Biloxi, MI.

APPENDIX B

REPRINTS AVAILABLE RELATED TO A MENU OF TESTS
FOR REPEATED STUDY OF HUMAN PERFORMANCE

REPRINTS AVAILABLE RELATED TO A MENU OF TESTS FOR REPEATED-MEASURES STUDY OF HUMAN PERFORMANCE

I.

Bittner, A. C., Jr., Smith, M. G., Kennedy, R. S., Staley, C. F., & Harbeson, M. M. (1985). Automated portable test system (APTS): Overview and prospects. Behavior Research Methods, Instruments and Computers, 17, 217-221.

Johnson, J. H., Kennedy, R. S., Smith, M. G., & Dutton, B. (1985). On the use of portable microprocessors as field data collection units. Proceedings of the Annual Scientific Meeting of the Aerospace Medical Association, San Antonio, TX.

Kennedy, R. S. (1985). A portable battery for objective, nonobtrusive measures of human performance. Proceedings of the Workshop on Advances in NASA-relevant, minimally invasive instrumentation (pp. 4.17-4.30). Pacific Grove, CA.

Kennedy, R. S. (1985). What are the advantages of self monitoring by the Automated Performance Test System (APTS)? Paper presented at the 38th International Air Safety Seminar, Boston, MA.

Kennedy, R. S., Dunlap, W. P., Jones, M. B., Lane, N. E., & Wilkes, R. L. (1985). Portable human assessment battery: stability, reliability, factor structure and correlation with tests of intelligence (Final Report NSF/BNS 85001; also EOTR 85-3). Washington, DC: National Science Foundation. NTIS PB88-116645/A03

Kennedy, R. S., Dunlap, W. P., Wilkes, R. L., & Lane, N. E. (1985). Development of a portable computerized performance test system. Proceedings of the 27th Annual Conference of the Military Testing Association (pp. 107-112). San Diego, CA: Navy Personnel Research & Development Center.

Kennedy, R. S., Jones, M. B., Dunlap, W. P., Wilkes, R. L., & Bittner, A. C., Jr. (1985). Automated Portable Test System (APTS): A performance assessment tool. SAE Technical Paper Series (Report No. 81775). Warrendale, PA: Society of Automotive Engineers. [wang 0139K]

Kennedy, R. S., Wilkes, R. L., Lane, N. E., & Homick, J. L. (1985). Preliminary evaluation of a microbased repeated measures testing system (EOTR 85-1; NASA CR-172038). Washington, DC: National Aeronautics and Space Administration. [wang 0056K]

Merkle, P. J., Kennedy, R. S., Smith, M. G., & Johnson, J. H. (1985). Microprocessor based field testing for human performance assessment. Proceedings of the 27th Annual Military Testing Association Conference (pp. 398-403). San Diego, CA: Navy Personnel Research & Development Center.

Dunlap, W. P., Bittner, A. C., Jr., Jones, M. B., & Kennedy, R. S. (1986, November). Factor analysis of composite scores from the Armed Services Vocational Aptitude Battery (ASVAB). Proceedings of the 28th Annual Military Testing Association Conference (pp. 218-224). Mystic, CT: U.S. Coast Guard Academy. [wang 0034J]

REPRINTS AVAILABLE RELATED TO A MENU OF TESTS FOR REPEATED-MEASURES STUDY OF HUMAN PERFORMANCE

Dunlap, W. P., Jones, M. B., Kemery, E. R., & Kennedy, R. S. (1986, November). Optimizing a test battery by varying subtest times. Proceedings of the 28th Annual Conference of the Military Testing Association (pp. 225-230). Mystic, CT: U.S. Coast Guard Academy. [wang 0429K]

Jones, M. B., Kennedy, R. S., & Turnage, J. J. (1986, September). Isoperformance: A methodology for human factors engineering design. Proceedings of the 30th Annual Meeting of the Human Factors Society, Dayton, OH. [wang 0254K]

Kennedy, R. S. (1986, October). A menu of performance tests implemented on a portable microcomputer. Proceedings of the 1st Joint IUTOX-IST Symposium on Behavioral Toxicology (to appear in the Journal of Neurobehavioral Toxicology and Teratology), Bari, ITALY. [wang 1424K]

Kennedy, R. S., & Kuntz, L. A. (1986). Self-monitoring of subjective status during extended operations using an Automated Performance Test Battery. Paper presented at the 37th International Astronautical Congress, Innsbruck, AUSTRIA. [wang 0356K]

Kennedy, R. S., Dunlap, W. P., & Kuntz, L. A. (1986). Application of a portable automated performance test battery for the study of drugs and driving performance. Paper presented at the 2nd International Symposium on Medicinal Drugs and Driving Performance. [wang 0030M]

Kennedy, R. S., Lane, N. E., Wilkes, R. L., & Banderet, L. E. (1986, November). Development of behavioral assessment protocols for varied repeated-measures testing paradigms. Proceedings of the 28th Annual Military Testing Association Conference (pp. 568-573). Mystic, CT: U.S. Coast Guard Academy. [wang 0033j]

*Kennedy, R. S., Wilkes, R. L., & Kuntz, L. A. (1986). Sensitivity of a notebook-sized Portable Automated Performance Test System. Paper presented at the Annual Behavioral Toxicology Society Meeting, Atlanta, GA.

Lane, N. E., Kennedy, R. S., & Jones, M. B. (1986). Overcoming unreliability in operational measures: The use of surrogate measure systems. Proceedings of the 30th Annual Meeting of the Human Factors Society (pp. 1398-1402). Dayton, OH: Human Factors Society. [wang 0016H]

Wilkes, R. L., Kennedy, R. S., Dunlap, W. P., & Jones, M. B. (1986). Development of a portable human assessment battery for environmental and behavioral toxicology studies. Unpublished manuscript. [wang 0019j]

Wilkes, R. L., Kennedy, R. S., Dunlap, W. P., & Lane, N. E. (1986). Stability, reliability, and cross-mode correlations of tests in a recommended 8-minute performance assessment battery (TR No. EOTR 86-4 for NASA Contract No. NAS9-17326). Essex Corporation, Orlando, FL.

Fowlkes, J. E., Kennedy, R. S., & Hennessy, R. T. (1987, August). Relevance of visual accommodation for performance in spacecraft (NASA Contract NAS9-17745). Final Report submitted to NASA Lyndon B. Johnson Space Center, Houston, TX. [wang 1509K]

REPRINTS AVAILABLE RELATED TO A MENU OF TESTS FOR REPEATED-MEASURES STUDY OF HUMAN PERFORMANCE

Jones, M. B., Kennedy, R. S., & Kuntz, L. A. (1987, August). Isoperformance: Integrating personnel and training factors into equipment design. Paper presented at the 2nd International Conference on Human-Computer Interaction, Honolulu, HI.

Jones, M. B., Kennedy, R. S., Kuntz, L. A., & Baltzley, D. R. (1987, October). Isoperformance: Trading off selection, training, and equipment variations to maintain the same level of systems performance. Proceedings of the 31st Annual Meeting of the Human Factors Society (pp. 634-637). Santa Monica, CA: Human Factors Society. [wang 1434K]

Kennedy, R. S., Berbaum, K. S., Williams, M. C., Brannan, J., & Welch, R. B. (1987). Transfer of perceptual-motor training and the space adaptation syndrome. Aviation, Space, and Environmental Medicine, 58(9,Suppl.), A29-A33.

Kennedy, R. S., Lane, N. E., & Kuntz, L. A. (1987, August). Surrogate measures: A proposed alternative in human factors assessment of operational measures of performance. Proceedings of the 1st Annual Workshop on Space Operations, Automation, & Robotics (pp. 551-558). Houston, TX: Lyndon B. Johnson Space Center. [wang 1487K]

Kennedy, R. S., Wilkes, R. L., Dunlap, W. P., & Kuntz, L. A. (1987). Development of an Automated Performance Test System for environmental and behavioral toxicology studies. Perceptual and Motor Skills, 65, 947-962.

Kennedy, R. S., Wilkes, R. L., Dunlap, W. P., & Kuntz, L. A. (1987, October). Microbased repeated-measures performance testing and general intelligence. Paper presented at the 29th Annual Conference of the Military Testing Association, Ottawa, Ontario, Canada. [wang 1570K]

Tabler, R. E., Turnage, J. J., & Kennedy, R. S. (1987, September). Repeated-measures analyses of selected psychomotor tests from PAB and APTS: Stability, reliability, and cross-task correlations (DAMD 17-85-C-5095). U.S. Army Aeromedical Research Laboratory, Fort Rucker, AL. [wang 1527K]

Turnage, J. J., Kennedy, R. S., & Osteen, M. K. (1987, September). Repeated-measures analyses of selected psychomotor tests from PAB and APTS: Stability, reliability, and cross-task correlations (DAMD 17-85-C-5095). U.S. Army Aeromedical Research Laboratory, Fort Rucker, AL. [wang 1470K]

Turnage, J. J., & Lane, N. E. (1987, October). The use of surrogate techniques for the measurement of team performance. Proceedings of the 31st Annual Meeting of the Human Factors Society (pp. 638-642). Santa Monica, CA: Human Factors Society. [wang 0260L]

Wilkes, R. L., Kuntz, L. A., & Kennedy, R. S. (1987, October). Development of a menu of performance tests self-administered on a portable microcomputer. NASA Technical Report under Contract NASA 9-17326. [wang 0193W]

Dunlap, W. P., Kennedy, R. S., Harbeson, M. M., & Fowlkes, J. E. (1988). Difficulties with individual difference measures based upon some componential cognitive paradigms. Manuscript submitted for publication. [wang 0073J]

REPRINTS AVAILABLE RELATED TO A MENU OF TESTS FOR REPEATED-MEASURES STUDY OF HUMAN PERFORMANCE

Fowlkes, J. E., Kennedy, R. S., Dunlap, W. P., & Harbeson, M. M. (1988, October). A paradigm for the identification of independent cognitive constructs. Paper presented at the 32nd Annual Meeting of the Human Factors Society, Anaheim, CA.

Jones, M. B., Kennedy, R. S., & Baltzley, D. R. (1988). Factor analysis of a microcomputer-based performance battery and its utility in predicting the Wonderlic Personnel Test (EOTR-88-8). Orlando, FL: Essex Corporation.

Kennedy, R. S., Baltzley, D. R., Osteen, M. K., & Turnage, J. J. (1988, October). A differential approach to microcomputer test battery development and implementation. Paper presented at the 32nd Annual Meeting of the Human Factors Society, Anaheim, CA.

Kennedy, R. S., Dunlap, W. P., Banderet, L. E., Smith, M. G., & Houston, C. S. (1988). Cognitive performance decrements occasioned by a simulated climb of Mount Everest: Operation Everest II. Unpublished manuscript. [wang 0907k]

Kennedy, R. S., Jones, M. B., & Baltzley, D. R. (1988, February). Empirical demonstration of Isoperformance methodology preparatory to development of an interactive expert computerized decision aid. Final Report submitted to Army Research Institute, Washington, DC. [wang 1646K]

Kennedy, R. S., Turnage, J. J., & Lane, N. E. (1988, June). Application of a portable microcomputer mental acuity battery for fitness-for-duty assessment in power plant operations. Paper presented at the IEEE 4th Conference on Human Factors and Power Plants, Monterey, CA. [wang 1633K]

Kennedy, R. S., Turnage, J. J., & Lane, N. E. (1988). Assessment of fitness-for-duty in power plant operations by a portable microcomputer mental acuity battery. Transactions of the American Nuclear Society, 56, 521-522. (ISSN:000-018X) [wang 1658K]

Kennedy, R. S., Wood, C., Baltzley, D. R., Odenheimer, R. S., & Dunlap, W. P. (1988). Effects of scopolomine and amphetamine on microcomputer performance tests. Orlando, FL: Essex Corporation.

Lane, N. E., & Kennedy, R. S. (Eds.). (1988, May). Users manual for the U.S. Army Aeromedical Research Laboratory Portable Performance Assessment Battery (Tech. Rep. No. EOTR 88-5). Ft. Rucker, AL: U.S. Army Aeromedical Laboratory.

Turnage, J. J., Kennedy, R. S., Osteen, M. K., & Tabler, R. E. (1988). Repeated-measures analyses of selected psychomotor tests from PAB and APTS: Stability, reliability, and cross-task correlations: Study 3. Orlando, FL: Essex Corporation. [wang 0045c]

Wilkes, R. L., Kuntz, L. A., Kennedy, R. S., & Tabler, R. E. (1988). Stability and reliability of a menu of performance tests self-administered on a portable microcomputer. NASA Contract NAS9-17326. NASA Technical Report. Houston, TX: NASA Johnson Space Center.

REPRINTS AVAILABLE RELATED TO A MENU OF TESTS FOR REPEATED-MEASURES STUDY OF HUMAN PERFORMANCE

Williams, M., Kennedy, R. S., Baltzley, D. R., May. J. G., & Dunlap, W. P. (1988). Reliability, stability, and cross-task correlations of six visual temporal factor tests. Essex Orlando Technical Report. Orlando, FL: Essex Corporation. [wang 1689K]

II. HISTORY - PETER, ETC.

Kennedy, R. S., & Bittner, A. C., Jr. (1977). The development of a Navy Performance Evaluation Test for Environmental Research (PETER). In L. T. Pope & D. Meister (Eds.), Productivity enhancement: Personnel performance assessment in Navy systems (pp. 393-408). Naval Personnel Research and Development Center, San Diego, CA. (NTIS No. AD A056047)

Kennedy, R. S., & Bittner, A. C., Jr. (1978). The stability of complex human performance for extended periods: Application for studies of environmental stress. Proceedings of the 49th Annual Meeting of the Aerospace Medical Association, New Orleans, LA.

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). Detroit, MI. (NTIS No. AD A060676)

Kennedy, R. S., & Bittner, A. C., Jr. (1978). Progress in the development of a Performance Evaluation Test for Environmental Research (PETER). Paper presented at the First Informal ONR Contractors' Meeting on Individual Differences in Cognitive Performance, Stanford University.

Damos, D. L., Kennedy, R. S., & Bittner, A. C., Jr. (1979). Development of Performance Evaluation Tests for Environmental Research (PETER): 2. Critical tracking test. Proceedings of the 50th Annual Meeting of the Aerospace Medical Association (pp. 33-34). Washington, DC. (NTIS No. AD A066719)

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.

Kennedy, R. S., Bittner, A. C., Jr., & Harbeson, M. M. (1979). Research developments in Performance Evaluation Tests for Environmental Research (PETER). Paper presented at the Annual Scientific Meeting of the Undersea Medical Society, Key Biscayne, FL. (Abstract in Undersea Biomedical Research, 1979, 6 (1, Supplement), 44.

Kennedy, R. S., Bittner, A. C., Jr., & Harbeson, M. M. (1979). The Performance Evaluation Tests for Environmental Research (PETER) Paradigm. Paper presented at the 20th Annual Meeting of the Psychonomic Society, Phoenix, AZ. (Abstract in Bulletin of the Psychonomic Society, 1979, 14, 248.)

Kennedy, R. S., Bittner, A. C., Jr., & Jones, M. B. (1979). Development of Performance Evaluation Tests for Environmental Research (PETER): Air Combat Maneuvering Test (ATARI). Paper presented at the Rocky Mountain Psychological Association Meeting, Las Vegas, NV.

*Bittner, A. C., Jr., Kennedy, R. S., & Harbeson, M. M. (1980). Apparatus testing for aviation performance assessment and selection: A technology ready to come of age. Paper presented at the 28th International Congress of Aviation and Space Medicine, Montreal, Quebec, Canada.

REPRINTS AVAILABLE RELATED TO A MENU OF TESTS FOR REPEATED-MEASURES STUDY OF HUMAN PERFORMANCE

Carter, R. C. Kennedy, R. S., Bittner, A. C., Jr., & Krause, M. (1980). Item recognition as a Performance Evaluation Test for Environmental Research (PETER). Proceedings of the 24th Annual Meeting of the Human Factors Society (pp. 340-344). Los Angeles, CA.

Carter, R. C., & Kennedy, R. S. (1980). Selection of Performance Evaluation Tests for Environmental Research (PETER). Proceedings of the 24th Annual Meeting of the Human Factors Society (pp. 320-324). Los Angeles, CA.

Bittner, A. C., Jr., Jones, M. B., Carter, R. C., Shannon, R. H., Chatfield, D. C., & Kennedy, R. S. (1981). Statistical issues in performance testing: Collected papers (NBDL 81R010). New Orleans, LA: Naval Biodynamics Laboratory. (NTIS No. AD A111086)

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

Damos, D. L., Bittner, A. C., Jr., Kennedy, R. S., & Harbeson, M. M. (1981). The effects of extended practice on dual-task training. Human Factors, 23, 627-631.

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., Andrews, D. A., & Carter, R. C. (1981). Performance Evaluation Tests for Environmental Research (PETER): A microcomputer game as a memory test. Proceedings of the 52nd Annual Scientific Meeting of the Aerospace Medical Association (pp. 240-241). San Antonio, TX.

Kennedy, R. S., Bittner, A. C., Jr., Carter, R. C., Krause, M., Harbeson, M.M., McCafferty, D. B., Pepper, R. L., & Wiker, S. F. (1981). Performance Evaluation Tests for Environmental Research (PETER): Collected papers (NBDL 80R008). New Orleans, LA: Naval Biodynamics Laboratory.

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

Harbeson, M. M., Kennedy, R. S., Krause, M., & Bittner, A. C., Jr. (1982). Repeated measures of information processing. Proceedings of the 26th Annual Meeting of the Human Factors Society (pp. 818-822). Seattle, WA: Human Factors Society.

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

Kennedy, R. S., Bittner, A. C., Jr., Harbeson, M. M., & Jones, M. B. (1982). Television computer games: A "new look" in performance testing. Aviation, Space, & Environmental Medicine, 53, 49-53.

REPRINTS AVAILABLE RELATED TO A MENU OF TESTS FOR REPEATED-MEASURES STUDY OF HUMAN PERFORMANCE

Harbeson, M. M., Kennedy, R. S., & Bittner, A. C., Jr. (1980). Application of memory tests for assessment of the effects of exotic environments on humans. Paper presented at the 72nd Annual Meeting of the Southern Society for Philosophy & Psychology, Birmingham, AL.

Harbeson, M. M., Krause, M., & Kennedy, R. S. (1980). The comparison of memory tests for environmental research. Proceedings of the 24th Annual Meeting of the Human Factors Society (pp. 349-353). Los Angeles, CA.

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

Kennedy, R. S., Bittner, A. C., Jr., & Harbeson, M. M. (1980). An engineering approach to the standardization of Performance Evaluation Tests for Environmental Research (PETER). Proceedings of the 11th Annual Conference of the Environmental Design and Research Association (EDRA), Charleston, SC.

Kennedy, R. S., Carter, R. C., & Bittner, A. C., Jr. (1980). A catalogue of Performance Evaluation Tests for Environmental Research (PETER). Proceedings of the 24th Annual Meeting the Human Factors Society (pp.344-348). Los Angeles, CA.

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 (pp. 108-113). Point Ideal, Ontario, Canada.

Krause, M., & Kennedy, R. S. (1980). Performance Evaluation Tests for Environmental Research (PETER): Interference susceptibility test. Proceedings of the 7th Psychology in the DoD Symposium (pp. 459-464). Colorado Springs, CO: USAF Academy.

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

Pepper, R. L., Kennedy, R. S., & Bittner, A. C., Jr. (1980). Development of Performance Evaluation Tests for Environmental Research (PETER): Code substitution test. Proceedings of the 7th Psychology in the DoD Symposium (pp. 451-458). Colorado Springs, CO: USAF Academy.

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

Bittner, A. C., Jr., Kennedy, R. S., & McCauley, M. E. (1980). Time estimation: Repeated-measures testing and drug effects. Proceedings of the 7th Psychology in the DoD Symposium (pp. 445-459). Colorado Springs, CO: USAF Academy.

REPRINTS AVAILABLE RELATED TO A MENU OF TESTS FOR REPEATED-MEASURES STUDY OF HUMAN PERFORMANCE

Bittner, A. C., Jr., Carter, R. C., Krause, M., Kennedy, R. S., & Harbeson, M. M. (1983). Performance Evaluation Tests for Environmental Research (PETER): Moran and computer batteries. Aviation, Space, and Environmental Medicine, 54, 923-928.

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

McCormick, B. K., Dunlap, W. P., Kennedy, R. S., & Jones, M. B. (1983). The effects of practice on the Armed Services Vocational Aptitude Battery (Report No. TR-602). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (NTIS No. AD A148314)

Smith, M. G., Krause, M., Kennedy, R. S., Bittner, A. C., Jr., & Harbeson, M. M. (1983). Performance testing with microprocessors--Mechanization is not implementation. Proceedings of the 27th Annual Meeting of the Human Factors Society (pp. 674-678). Norfolk, VA: Human Factors Society.

Wiker, S. F., Kennedy, R. S., & Pepper, R. L. (1983). Performance Evaluation Tests for Environmental Research (PETER): Navigational plotting task. Aviation, Space, and Environmental Medicine, 54, 144-149.

Banderet, L. E., Benson, K. P., McDougall, D. M., Kennedy, R. S., & Smith, M. (1984). Development of cognitive tests for repeated performance assessment. 1Proceedings of the 26th MTA Conference (pp. 375-380), Munich, Germany.

Bittner, A. C., Jr., Carter, R. C., Kennedy, R. S., Harbeson, M. M., & Krause, M. (1984). Performance Evaluation Tests for Environmental Research (PETER): Evaluation of 112 measures (Research Report NBDL-84R006). New Orleans, LA: Naval Biodynamics Laboratory. (NTIS No. AD A152317)

Damos, D. L., Bittner, A. C., Jr., Kennedy, R. S., Harbeson, M. M., & Krause, M. (1984). Performance Evaluation Tests for Environmental Research (PETER): Critical tracking test. Perceptual & Motor Skills, 58, 567-573.

*Kennedy, R. S., Bittner, A. C., Jr., Smith, M. G., & Harbeson, M. M. (1984). Development of a portable performance assessment system for behavioral toxicology. Paper presented at the Behavioral Toxicology Society Meeting, Toronto, Canada.

Lintern, G., & Kennedy, R. S. (1984). A video game as a covariate for carrier landing research. Perceptual & Motor Skills, 58, 167-172.

Pepper, R. L., Kennedy, R. S., Bittner, A. C., Jr., Wiker, S. F., & Harbeson, M. M. (1985). Performance Evaluation Tests for Environmental Research (PETER): Code substitution test. Perceptual and Motor Skills, 61, 735-745.

Bittner, A. C., Jr., Carter, R. C., Kennedy, R. S., Harbeson, M. M., & Krause, M. (1986). Performance Evaluation Tests for Environmental Research (PETER): Evaluation of 114 measures. Perceptual and Motor Skills, 63, 683-708.

REPRINTS AVAILABLE RELATED TO A MENU OF TESTS FOR REPEATED-MEASURES STUDY OF HUMAN PERFORMANCE

III. IN PREPARATION

Baltzley, D. R., Dunlap, W. P., Kennedy, R. S., Odenheimer, R. S., & Wood, C. (1988). Microcomputer based performance testing: The effects of scopolamine and amphetamine. Orlando, FL: Essex Corporation.

Ginsburg, A. P., Kennedy, R. S., Wilkes, R. L., & Baltzley, D. R. (1988). Assessment of vision measures for repeated-measures applications: The VISTECH Contrast Sensitivity Test (National Science Foundation Contract BNS-8460765).

Jones, M. B., Kennedy, R. S., & Turnage, J. J. (1988). Factor and regression analyses of selected tests from the APTS and PAB batteries (NSF Grant ISI-8521282). Final Report. National Science Foundation, Washington, D.C.

Kennedy, R. S., Jones, M. B., Baltzley, D. R., & Turnage, J. J. (1988). The prediction of intelligence and the factor structure of selected tests from two performance assessment batteries. Funding provided by the National Science Foundation, Washington, D.C. under NSF Grant ISI-8521282. Unpublished manuscript. [wang 1637K]

Kennedy, R. S., & Turnage, J. J. (1988). Repeated-measures analyses of selected psychomotor tests from PAB and APTS: Stability, reliability, and cross-task correlations. Orlando, FL: Essex Corporation.

Parth, P., Lane, N. E., Dunlap, W. P., Chapman, R., Kennedy, R. S., & Ordy, J. M. (1988). Cognitive deficits resulting from chemoradiotherapy in bone marrow transplant patients. Unpublished manuscript.