CR-185518

A MENU OF SELF-ADMINISTERED MICROCOMPUTER-BASED NEUROTOXICOLOGY TESTS

Robert S. Kennedy, Robert L. Wilkes,
Lois-Ann Kuntz, and Dennis R. Baltzley

Essex Corporation
1040 Woodcock Road, Suite 227
Orlando, FL 32826

EOTR 88-10

November 1988

ACKNOWLEDGMENTS

This effort was performed under Contract No. NAS9-17326 for National Aeronautics and Space Administration, Lyndon B. Johnson Space Center, Space and Life Sciences Procurement Office, Houston, TX. The authors are indebted to Mary K. Osteen for editorial assistance.

ABSTRACT

This study examined the feasibility of repeated self-administration of a newly developed battery of mental acuity tests. We developed this battery to be used to screen the fitness for duty of persons in at-risk occupations (astronauts, race car drivers), or those who may be exposed to environmental stress, toxic agents, or disease. The menu under study contained cognitive and motor tests implemented on a portable microcomputer including: a five-test "core" battery, lasting six minutes, which had demonstrable reliabilities and stability from several previous repeated-measures studies, and also 13 "new" tests, lasting 42 minutes, which had appeared in other batteries but had not yet been evaluated for repeated-measures implementation in this medium.

Sixteen subjects self-administered the battery over 10 repeated sessions. The hardware performed well throughout the study and the tests appeared to be easily self-administered. Stabilities and reliabilities of the tests from the core battery were comparable to those obtained previously under more controlled experimental conditions. Analyses of metric properties of the remaining 13 tests produced eight additional tests with satisfactory properties. Although the average retest reliability was high, cross-correlations between tests were low, indicating factorial richness. The menu can be used to form batteries of flexible total testing time which are likely to tap different mental processes and functions.

INTRODUCTION

PREMISE

The presence of environmental stressors and toxic elements in space exploration, the military, and the workplace makes desirable the development of an assessment tool to detect subtle differences in mental acuity, performance, and health. The tests could also be used for monitoring the neurological status of persons subjected to hazards in their occupations, such as deep sea divers or boxers, as well as for longitudinal monitoring in connection with regular physical examinations.

To be effective as an on-line screening instrument, the battery should be easily administered at different sites and have many equivalent alternate forms. To be most effective the instrument should employ objective tests of complex mental functioning which could be self-administered by the person who is exposed rather than requiring a trained proctor.

BACKGROUND

The lack of a standardized, sensitive human performance assessment battery has probably delayed recognition of the deleterious effects of marijuana (Nicholi, 1983; Turner, 1983) and is recognized as a particularly important need in the field of behavioral toxicology. "In general, there is an exclusion of behavior from food additive testing protocols...although one of the reasons for its exclusion is a lack of confidence in currently proposed behavioral tests" (Weiss, 1983, p. 1185). The Toxic Substances Control Act of 1976 specifies behavior (Michael, 1982) as one of the criteria for judging the safety of new chemicals, but probably no satisfactory battery is available should the requirement be applied. One which shows promise must be proctored by trained administrators in a laboratory setting (Hanninen & Landstrom, 1979). Studies of toxic waste, side effects of drugs, industrial exposure of potentially hazardous materials, food additives, over-the-counter pharmaceuticals, controlled substances, alcohol, dietary supplements, and exposures to other chemical substances all require a performance test battery to address possible subtle behavioral impacts. In addition, with the availability of such a test battery studies of the performance effects of environmental stressors of interest to the military and others could be conducted such as thermal extremes, hyper- or hypobaria, motion, vibration, noise, sensory deprivation or overload. Other applications include study of conditions or processes such as aging, dementia, sleep deprivation or emotional strain.

If successful, this testing tool could be used to screen key persons in responsible jobs (e.g., nuclear power plants) for fitness for duty, and to predict premonitory onset of decrements in performance, physiology, mood, and behavior before such changes threaten operational efficiency. Such a battery could be used to provide feedback to susceptible personnel, to explore the possibility of coping methods, adaptation and resistance training, and to monitor the neurological status of persons subjected to hazards (astronauts, deep sea divers) and risks (race car drivers) in their occupations, as well as for longitudinal monitoring in connection with regular physical examinations.

High-speed portable personal computers are now widely available for repeated-measures performance testing. These devices can have several obvious advantages. Increased control and standardization of testing conditions by use of a computerized battery of tests should lead to: a) more accurate and objective response scoring, b) the elimination of clerical errors in data transfer and subjective interpretation, c) utilization of response latency measures, and d) higher test reliabilities. Early in the development of tasks (and batteries), concerns were raised about whether implementation in a new medium would change what the test tested and how well, and so much of our early work addressed this issue (Barrett, Alexander, Dovberspike, Cellar, & Thomas, 1982; Kennedy, Wilkes, Lane, & Homick, 1985; Smith, Krause, Kennedy, Bittner, & Harbeson, 1983).

The primary purpose of the present study was to continue with our development of a metrically sound human performance test battery suitable for repeated-measures research by evaluating tests of other factors and comparing them to our core battery. Previous studies had surfaced stable paper-and-pencil tests with "good" metric properties (Bittner, Carter, Kennedy, Harbeson, & Krause, 1986) that were then mechanized on a portable microcomputer and are now available (Kennedy, Lane, & Kuntz, 1987). Recently, in order to guide future test development a task analysis (Jeanneret, 1988) of the activities of space travelers (astronauts and payload specialists) was performed and used to evaluate the tests of the "core" battery thus far available. The tests selected for the present study were included to add constructs which appeared to be missing from the core battery.

Eighteen microbased tests (13 new, 5 core) were examined. Among other tests, this new version contained reaction time and more complex visual and auditory short-term memory tests. A second, but equally important, purpose was to assess the viability of subject self-administration of the battery in nonlaboratory environments. We were therefore anxious to determine whether the test battery was sufficiently "friendly" that it could be self-administered under field conditions degrading reliability and predictive validity.

METHOD

SUBJECTS

Eighteen freshman and sophomore students from the University of Wyoming and Casper College at Casper, Wyoming, participated in the study. The individuals were solicited from a pool of subjects from psychology classes. Subject procurement and data collection procedures were carried out in accordance with APA principles for research with human subjects (American Psychological Association, 1982). Subject motivation for participation was high with 100% of the contacted individuals volunteering, although one subject ended up being removed from the study for noncompliance with testing protocol. A second subject was lost because midway through the experiment his data were inadvertently destroyed during a transfer process, and he too was dropped. Final analyses were based on data obtained from the remaining subjects (nine women and seven men).

PROCEDURE

All testing was accomplished with a fully automated portable microcomputer system. The microbased battery of eighteen subtests was programmed to be self-administered over 10 sessions of testing. Prior to initial testing, subjects were thoroughly introduced to the purpose and nature of the study and pertinent biographical data were obtained. Special attention was given to subject training, orientation, and indoctrination during session 1. Testing schedules were established relative to the subject's personal needs. Tests were administered at most twice on any day over a 10-day period at times amenable to data collection. Departures were allowed within certain limitations, and the prevailing criterion being subject motivation. Self-administration of the first battery was completed in the experimenter's presence to ensure knowledge of system operation and to surface questions.

Special efforts were made to ensure that each subject understood the consequences to the study of engaging in activities likely to influence test performance in adverse and uncontrolled ways. Subjects were informed that the performance tests were the focus of the study as opposed to the individuals themselves, and handouts and reminders concerning the test system operation and testing protocol were provided. The potential effects of drugs, alcohol, fatigue, emotional distress, illness, and other internal or environmental agents on behavior were reviewed and stressed. Subjects were directed not to test themselves if they believed, for any reason, their performance would be compromised. Whereas statistical power benefits greatly from replications, particularly when retest reliability is high (Dunlap, Jones, & Bittner, 1983), it is noted that repeated-measures studies are vulnerable to such effects, particularly if they are introduced systematically.

The microprocessor capability for monitoring test performance on a date/time basis was demonstrated and subjects were informed that testing would be checked prior to final payment. The microprocessors were "safed" to prevent memory access, and score tampering and there was no feedback or knowledge of results.

APPARATUS

Microcomputer testing was accomplished with the Automated Performance Test System (APTS) implemented on the NEC PC8201A microprocessor (Bittner, Smith, Kennedy, Staley, & Harbeson, 1985). The NEC PC8201A is configured around an 80C85 microprocessor with 64K internal ROM containing Basic, TELCOM, and a TEXT EDITOR. RAM capacity may be expanded to 96K onboard, divided into three separate 32K banks. An RS-232 interface allows for hook-up to modem, to a CRT or flat-panel display, to a "smart" graphics module, to a printer, or to other computer systems. Visual displays are presented on a 8-line LCD with 40 characters per line. Memory may be transferred to 32K modules with independent power supplies for storage or mailing. The entire package is lightweight (3.8 lbs), compact (110W x 40H x 130D mm), and fully portable with rechargeable nickel cadmium batteries permitting up to four hours of continuous operation. The technical features of the system which are more fully described in NEC Home Electronics (1983) and Essex Corporation (1985).

MATERIALS

The microbased test battery consisted of 18 individual performance subtests described below, which appear in Table 1 along with their administration times. (The tests are available on request on an IBM computer floppy disk from Dr. Robert S. Kennedy at Essex Corporation, 1040 Woodcock Road, Suite 227, Orlando, Florida, 32803.)

TABLE 1. MICROBASED BATTERY TASK ORDER AND TESTING TIME

Battery Task Order	Trials/Practice Admin.	Time	Total Task Time Each Administration	Total Task Time for 10 (incl. practice Administrations
1. Preferred Hand Tap*	2	10[a]	20	210
2. Reaction Time (1 Choice)	1	30	120	1230
3. Auditory Count (1 Stimulus)	1	0	300	3000
4. Short-term Memory	1	30	120	1230
5. Auditory Count (2 Stimuli)	1	0	300	3000
6. Number Comparison	1	30	45	480
7. Auditory Count (3 Stimuli)	1	0	300	3000
8. Air Combat Maneuv.	1	0	120	1200
9. Reaction Time (2 Choice)	1	30	120	1230
10. Two-Hand Tapping*	2	10	20	210
11. Pattern Comparison*	1	30	120	1230
12. Visual Count (1 Stimulus)	1	0	300	3000
13. Associative Memory	1	0	90	900
14. Visual Count (2 Stimuli)	1	0	300	3000
15. Grammatical Reason.*	1	30	120	1230
16. Reaction Time (4 Choice)	1	30	120	1230
17. Visual Count (3 Stimuli)	1	0	300	3000
18. Nonpref. Hand Tap.*	2	10	20	210
Totals		240	2835	28590

[a] All time data are reported in seconds
* Core Battery

CORE BATTERY

Tapping. The test is accomplished by alternately pressing keys on the microprocessor keyboard. The task was administered in three different forms: (a) Preferred-hand Tapping (PTAP); (b) Two-hand Tapping (THTAP); and (c) Nonpreferred-hand Tapping (NTAP). Performance is based on the number of alternate key presses made in the allotted time (Kennedy, Wilkes, Lane, & Hommick, 1985).

Pattern Comparison (PC). The Pattern Comparison task (Klein & Armitage, 1979) is accomplished by the subject examining a pair of dot patterns and determining whether they are similar or different.

Grammatical Reasoning (GR). The Grammatical Reasoning Test (Baddeley, 1968) involves five grammatical transformations on statements about the relationship between two letters A and B. The five transformations are: (1) active versus passive construction, (2) true versus false statements, (3) affirmative versus negative phrasing, (4) use of the verb "precedes" versus the verb "follows," and (5) A versus B mentioned first. There are 32 possible items arranged in random order. The subject's task is to respond "true" or "false," depending on the verity of each statement.

NEW TESTS

Number Comparison (NC). The Number Comparison task (Ekstrom, French, Harman, & Dermen, 1976) involves the presentation and comparison of two sets of numbers. The subject's task is to compare the first and second set and decide if they are the same or different.

Short-term Memory (STM). The Short-term Memory Task (Sternberg, 1966) involves the presentation of a set of four digits for one second (positive set), followed by a series of single digits presented for two seconds (probe digits). The subject's task is to determine if the probe digits accurately represent the positive set and respond with the appropriate key press. Performance is based on the number of probes correctly identified.

Air Combat Maneuvering (ACM). The Air Combat Maneuvering test emulates a combat-type video game. The subject's task is to "shoot" a randomly moving stimulus target. The subject laterally positions and fires a projectile through activation of appropriate microprocessor keys.

Reaction Time. In this version, on each session the visual stimulus is prefaced by a variably timed auditory signal. The task was administered in three different forms: (a) 1-Choice (RT1), (b) 2-Choice (RT2), and (c) 4-Choice (RT4). Reaction time is measured from the onset of the visual stimulus to the key press (Donders, 1968).

Counting (Auditory and Visual). The Counting tests (Jerison, 1955; Kennedy & Bittner, 1980) are accomplished by the subject accurately monitoring the repeated occurrence of either a visual or auditory stimulus. The subject must indicate when a stimulus has been presented four times in succession and then repeat the monitoring process until the end of the session. The complexity (i.e., task loading) of the task may be altered by presenting one,

two, or three stimuli during the same session and requiring the subject to monitor each. In the auditory test mode, the stimuli were varied by presenting "beeps" of three different frequencies, and in the visual task mode, the stimuli were varied by presenting lighted boxes at different locations on the screen. For the low demand situations one stimulus was presented (low tone for Auditory Counting and the right side of the screen for Visual Counting); for the medium demand two stimuli were presented (low and high tones for the Auditory Counting and right and left side of the screen for Visual Counting); and for the high demand three stimuli were presented (low, middle, and high tones for Auditory Counting and right, middle, and left sides of the screen for Visual Counting).

Associative Memory (AM). This is a memory test (Underwood, Boruch, & Malmi, 1977) which requires the participant to view five sets of three letters that are numbered 1 to 5 and then to 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.

ANALYSIS AND SCORING

Although there are obvious advantages associated with self-administered automated computerized testing, there are also problems. In particular, because the data are analyzed remotely in time and space, it is necessary to specify which of many possible scores are acceptable, or conversely to screen for anomalies after the fact such as reaction times which are too short, percent correct scores of 50% which indicate random responding, etc., and to facilitate the selection of appropriate and representative scores for analyses. In the present study, while the computer and software were considered to perform in a very creditable manner, data anomalies were surfaced by graphing performances for clusters of three to five subjects for all 10 sessions of each test. As a result of these comparisons, the following problems and corrections were identified: (a) a programming error in the Grammatical Reasoning test for some subjects' sessions required that the number correct score be discarded (and thereby percent correct also); experience (Turnage, Kennedy, & Osteen, 1987) and logic imply this was not a critical loss since, when sessions are of fixed length, hits and latencies are often simple transforms of each other. (b) a second programming error resulted in the nonadministration of the Nonpreferred-hand Tapping task to the two left-handed subjects. As a result of the omission, no data on that test for those two subjects were entered; and (c) atypical scores were observed for each subject on the first session of Number Comparison in Session 1, which has subsequently been traced to a software error. Those scores were not analyzed. The programming errors have been subsequently identified and corrected.

The computer programs are designed to output number of items administered, number correct, number wrong, and response latency for most tasks. From these options most traditional scorings are possible. Our general philosophy of scoring is to use the method with the highest reliability provided it is also rational. Generally, this is number correct (Turnage, Kennedy, & Osteen, 1987). For some tests, latency (e.g., reaction time) is preferred and often is as good as number correct. Almost always in previous studies "right minus wrong" has been as good as "hits" and "latency," but

generally adds no new information. Percent correct is a derived score (Cronbach & Furby, 1970), and while the most commonly found in the scientific literature, is invariably less reliable (Seales, Kennedy, & Bittner, 1980; Turnage et al., 1987) and so may lack statistical power. However, percent correct or other derived scores should not always be avoided and we have made exceptions (Kennedy, Dunlap, Bandaret, Smith, Houston, 1988) to our advice against their use. With occasional exceptions (e.g., log latency for strings of reaction times) other scores are almost always poorer and not used. In the present study 26 "rational" scores were used in preliminary analyses. Afterwards, one score per test was selected for presentation of the findings.

Repeated-Measures Assessment and Selection Criteria

Following data inspection, each subtest was evaluated relative to repeated measures selection criteria. These criteria have been previously identified and discussed in the literature (Bittner, Carter, Kennedy, Harbeson & Krause, 1986; Jones, 1980) and are briefly reviewed below:

Stability. Repeated-measures studies of environmental influences on performance require stable measures if changes in the treatment (i.e., the environment) are to be meaningfully related to changes in performance (Jones, 1970a). Of particular concern is the fact that a subject's scores may differ significantly over time owing to instability of the measure. For example, the Jones two-process theory of skill acquisition (Jones, 1970a, 1970b) maintains that the advancement of a skill involves an acquisition phase in which persons improve at different rates, and a terminal phase, in which persons reach or approximate their individual limits. The theory further implies that when the terminal phase is reached, scores will cease to deviate, despite additional practice. Unless tests have been practiced to this point of differential stability, the determination of whether changes in scores are due to practice or some other variable is problematic. Therefore, a stable test implies that

forms of classical test theory (Allen & Yen, 1979). For example, in a study of the effects of a toxic substance, if scores on a performance test remained the same before or after exposure, and if the test were not differentially stable, it would not be possible to determine whether a decline in performance was masked by practice effects or whether there was no treatment effect. Only after differential stability is clearly and consistently established between subjects can the investigator place confidence in the adequacy of his measures.

In this study means were considered stable if they were level, asymptotic or showed zero rate of change of slope over sessions. Standard deviations were considered stable if constant over sessions. Correlations were evaluated by a new graphical method. First, the average correlation of each session with all other sessions was computed, i.e. the average correlation of each row of the correlation matrix excluding the diagonal element. This was compared to the "off diagonal average" defined as the average of the three correlations among a given session and the two following sessions, i.e. for the first stability point the average of r_{12}, r_{13}, and r_{23} are used. Stability was said to occur after that session where high ($r > .707$) and level cummulative average correlations were obtained. Additionally, the off diagonal average correlation plots should be parallel to the average

correlations of a trial with all other trials. Two examples of this method are shown in Figure 1 (stable correlations) and Figure 2 (unstable correlations).

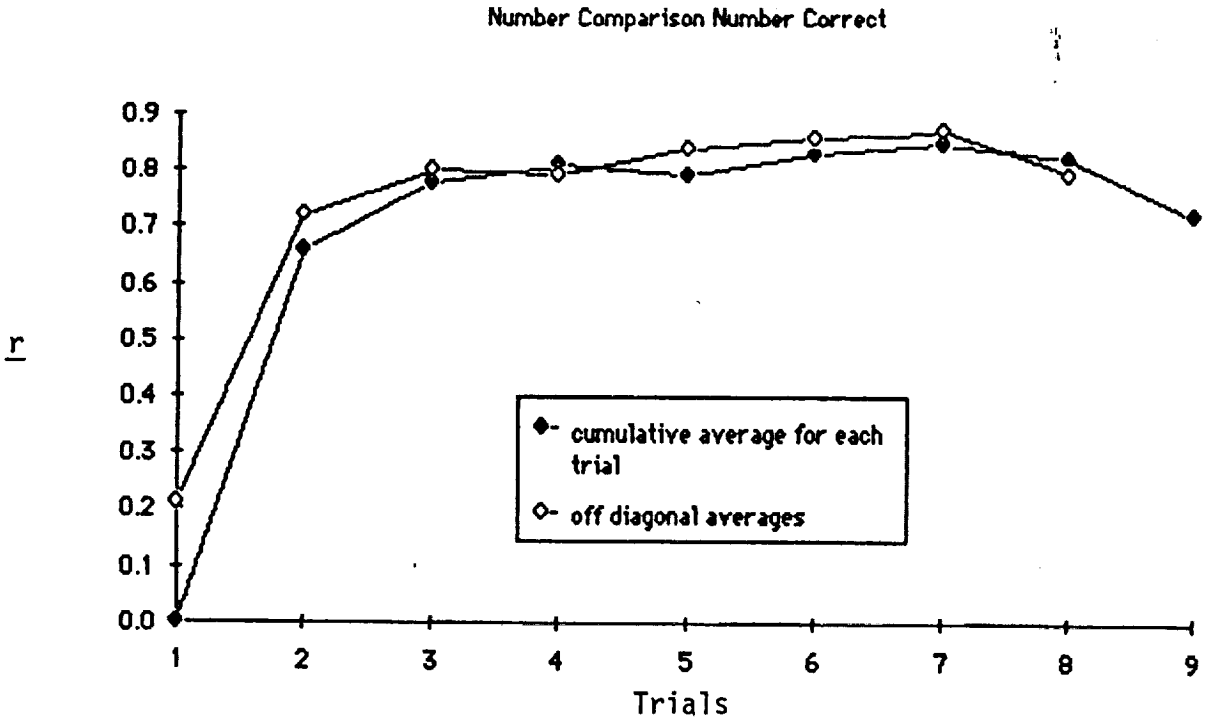

Figure 1. Correlational stability analysis for number comparison.

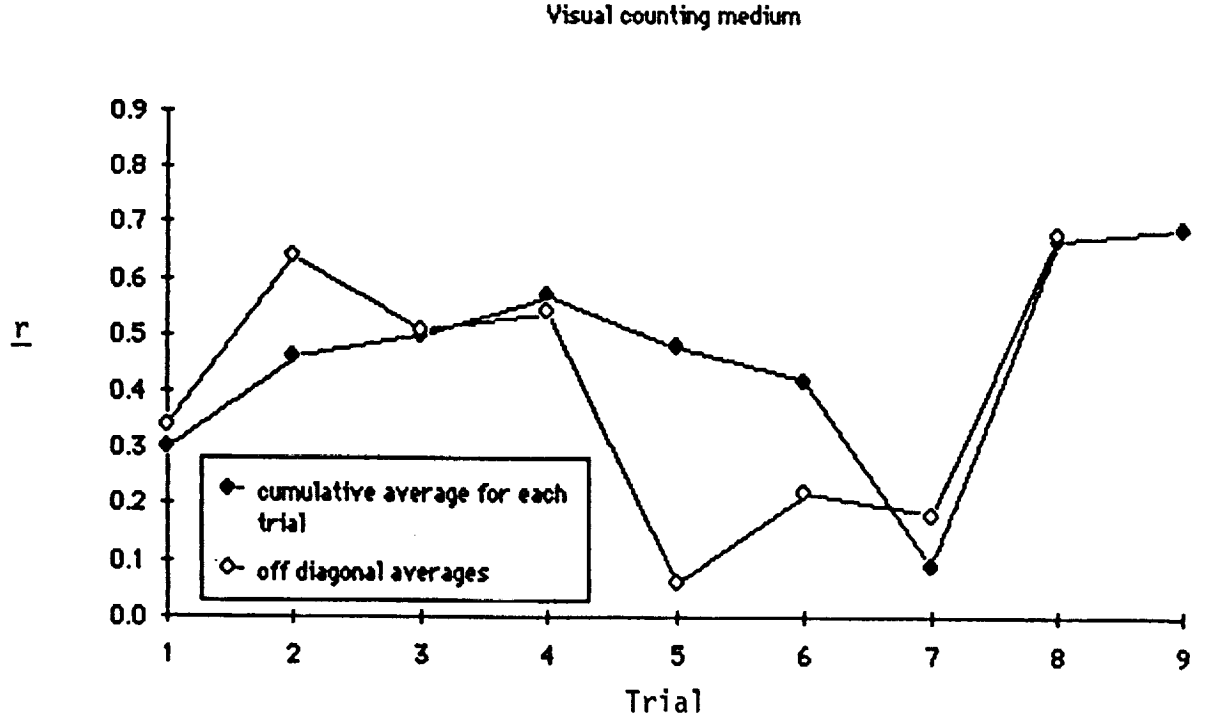

Figure 2. Correlational stability analysis for visual counting (medium difficulty).

Task Definition. Task Definition is the average reliability of the stabilized task (Jones, 1980). Task Definition is obtained by averaging stable intertrial correlations. Higher average reliability improves power in repeated-measures studies when variances are constant across sessions. The lower the error within a measure the greater the likelihood that mean differences will be detected, provided variances are also well behaved. Therefore, tasks with low task definition are insensitive to such differences and are to be avoided. Because different tasks stabilize at different levels, task definition becomes an important criterion in task selection. Task definitions for different tests, however, cannot be directly compared without first standardizing tests for test length (i.e., reliability efficiency).

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 tests with shorter administration times and/or fewer items. Test length must be equal before meaningful comparisons can be made. A useful tool for making relative judgments is the reliability-efficiency, or standardized reliability, of the test (Kennedy, Wilkes, Dunlap, & Kuntz, 1987). Reliability-efficiencies are computed by correcting the reliabilities of different tests to a common test length by use of the Spearman-Brown prophecy formula (Guilford, 1954, p. 354). Reliability-efficiency not only facilitates judgments concerning different tests, but also provides a means for comparing the sensitivity of one test with the sensitivity of another test.

Stabilization Time. The evaluation of highly transitory changes in performance may be necessary when studying the effects of various treatments, drugs, or environmental stress. Good performance measures should quickly stabilize following short periods of practice without sacrificing metric qualities, and good performance measures should always be economical in terms of time. A task under consideration for environmental research must be represented in terms of the number of sessions and/or the total amount of time necessary to establish stability. Stabilization time must be determined for the group means, standard deviations, and intertrial correlations (differential stability).

RESULTS

GENERAL

Group means and standard deviations were examined for evidence of test stabilization and intertrial correlations were assessed for evidence of correlational stability (i.e., differential stability), as well as task definitions and reliability efficiency. The means and standard deviations for the 18 tests appear in Table 2 where, over the 10 sessions, mean latencies (RL) appear to decrease, and mean hits (N) and number correct (NC) improve.

The findings of Table 2 are summarized in Table 3 by listing the day (session) of stability for means, standard deviations, and cross-session correlations. Figures 1 and 2 show descriptively examples of the correlational stability analyses. Those tests where correlations are level (e.g., Tapping, Short-Term Memory, Number Comparison) are stable from that session on (see Figure 1). Those which are not (e.g., Auditory Counting 1 &

2, Visual Counting 1 & 2) are clearly evident (see Figure 2). All standard deviations appeared stable but the four counting tests with the lowest workload demands (low and medium, auditory and visual) had poor correlational stability most likely due to the few opportunities for responding and so were judged unstable. The final column in Table 3 shows the overall trial of stability for each test using the latest trial from the three different stability measures.

TABLE 2. MEANS AND STANDARD DEVIATIONS

	Trials									
	1	2	3	4	5	6	7	8	9	10
Subtests										
1. PTAP(N)*	40	43	44	44	44	44	43	45	45	45
	(10)**	(10)	(9)	(10)	(10)	(9)	(9)	(10)	(9)	(10)
2. RT1(RL)	.33	.30	.28	.28	.27	.27	.27	.27	.29	.28
	(.57)	(.42)	(.36)	(.29)	(.43)	(.39)	(.40)	(.39)	(.81)	(.59)
3. ACT1(NC)	6	7	7	7	7	7	7	7	6	7
	(2)	(1)	(.6)	(.3)	(.7)	(.9)	(.8)	(1)	(1)	(.6)
4. STM(NC)	67	66	69	69	69	70	70	71	69	71
	(6)	(6)	(8)	(8)	(7)	(8)	(7)	(6)	(6)	(6)
5. ACT2(NC)	11	11	11	11	11	12	11	11	11	11
	(2)	(3)	(2)	(2)	(3)	(2)	(3)	(2)	(2)	(2)
6. NCP(NC)	NA	42	43	44	46	45	46	46	47	44
	NA	(7)	(8)	(9)	(9)	(9)	(10)	(9)	(8)	(13)
7. ACT3(NC)	13	14	14	14	15	15	14	15	13	15
	(4)	(4)	(3)	(4)	(3)	(3)	(4)	(3)	(5)	(3)
8. ACM(N)	78	89	94	100	104	104	103	110	112	110
	(17)	(24)	(24)	(22)	(20)	(16)	(19)	(17)	(20)	(17)
9. RT2(RL)	.44	.36	.34	.33	.33	.34	.32	.32	.33	.32
	(.23)	(.58)	(.51)	(.51)	(.57)	(.47)	(.40)	(.47)	(.46)	(.38)
10. THTAP(N)	45	46	47	46	47	47	47	46	47	48
	(10)	(11)	(11)	(10)	(10)	(10)	(11)	(10)	(12)	(12)
11. PC(NC)	86	89	92	93	97	97	97	99	98	99
	(13)	(12)	(13)	(12)	(13)	(14)	(13)	(14)	(13)	(11)
12. VCT1(NC)	7	7	7	7	7	7	7	7	7	7
	(.6)	(.5)	(.6)	(.4)	(.3)	(.5)	(.7)	(.8)	(.6)	(.9)
13. AM(NC)	11	12	13	13	14	15	14	14	15	15
	(4)	(3)	(5)	(5)	(4)	(4)	(3)	(5)	(5)	(4)
14. VCT2(NC)	13	12	12	12	12	12	13	12	12	12
	(.8)	(2)	(2)	(2)	(1)	(2)	(1)	(3)	(2)	(2)
15. GR(RL)	.31	.29	.18	.27	.28	.26	.26	.26	.26	.27
	(.87)	(.80)	(.89)	(.76)	(.92)	(.76)	(.77)	(.69)	(.87)	(.79)
16. RT4(RL)	.49	.44	.45	.42	.41	.41	.40	.42	.40	.40
	(.81)	(.95)	(.82)	(.64)	(.96)	(.76)	(.61)	(.74)	(.75)	(.67)
17. VCT3(NC)	16	16	16	16	15	16	17	16	16	16
	(4)	(4)	(3)	(3)	(3)	(3)	(2)	(2)	(3)	(3)
18. NTAP(N)	34	37	37	38	38	38	39	38	39	39
	(8)	(9)	(9)	(10)	(9)	(8)	(9)	(9)	(8)	(8)

* Codes: (N)=Number of Hits, (NC)=Number Correct, (RL)=Response Latency
** Standard Deviations in Parentheses
NA=Not analyzed due to software error or problems
Full test names are given by corresponding number in Tables 1 & 3.

TABLE 3. TRIAL AT WHICH STABILITY IS ACHIEVED

Variable	Mean	Standard Deviation	Correlation	Total Task Aggregation of Cols. 1,2,3
1. Preferred Hand Tapping (PTAP)	2	1	3	3
2. Average Reaction Time 1 (RT1)	3	6	3	6
3. Auditory Counting Low NC (ACT1)	2	2	UNST	UNST
4. Short Term Memory NC (STMNC)	3	3	1	3
5. Auditory Counting Med. NC (ACT2)	1	1	UNST	UNST
6. Number Comparison NC (NCNC)	3	3	3	3
7. Auditory Counting High NC (ACT3)	1	1	3	3
8. Air Combat Maneuvering (ACM)	UNST	5	UNST	UNST
9. Average Reaction Time 2 (RT2)	3	3	3	3
10. Two Hand Tapping (THTAP)	2	1	2	2
11. Pattern Comparison NC (PCNC)	5	1	3	5
12. Visual Counting Low NC (VCT1)	1	2	UNST	UNST
13. Associative Memory NC (AMNC)	3	3	5	5
14. Visual Counting Med. NC (VCT2)	1	2	UNST	UNST
15. Grammatical Reasoning RL (GRRL)	4	4	1	4
16. Average Reaction Time 4 (RT4)	2	1	2	2
17. Visual Counting High NC (VCT3)	1	1	6	6
18. Nonpreferred Tapping (NTAP)	2	2	1	2

TABLE 4. TASK DEFINITION (AVERAGE STABILIZED RELIABILITIES) FOR ALL TEST SCORES FOR ACTUAL TEST LENGTH AND ESTIMATED 3-MINUTE TEST LENGTH

Variable Minutes	Task Definition	Test Length(Sec)	Reliability for a 3-Minute Test
Preferred Hand Tap	.99	20	.99
Average Reaction Time 1	.59	120	.68
Auditory Counting Low NC	*	300	*
Short-Term Memory NC	.95	120	.97
Aud. Counting Medium NC	*	300	*
Number Comparison NC	.92	45	.98
Auditory Counting High NC	.78	300	.68
Air Combat Maneuvering	*	120	*
Average Reaction Time 2	.94	120	.96
Two Hand Tapping	.97	20	.99
Pattern Comparison NC	.94	120	.96
Visual Counting Low NC	*	300	*
Associative Memory NC	.80	90	.89
Visual Counting Medium NC	*	300	*
Grammatical Reasoning (RL)	.96	120	.97
Average Reaction Time 4	.94	120	.96
Visual Counting High NC	.78	300	.68
Nonpreferred Tapping	.98	20	.99

*Task definition can only be calculated meaningfully for stable tests.

Table 4 contains the obtained task definitions (stabilized retest reliability) as well as the predicted task definitions for a three minute test for the 13 tests which stabilized. The obtained task definitions are used later in Table 5 where the stabilized retest reliabilities appearing in the diagonal. The predicted value is derived from the substitution of "time in seconds" for "number of items" in the Spearman adjustment equation for test length (Guilford, 1954). Obtained task definitions ranged from r=.59 to .99 and after being normalized to the three minute base, most tests continued to remain suitable for repeated-measures usage according to our criteria. There were exceptions: one test which was marginally unacceptable became acceptable (Simple RT), and two other tests become nearly unacceptable (Auditory and Visual Counting High Demand). The range of normalized reliabilities in Table 4 varies from r=.68-.99 and except for the Counting and Simple Reaction Time tests all exhibited very high reliabilities (r > .89).

Summaries of the results for each test follows:

The Tapping Series. These tasks stabilized quickly and had high reliabilities for each of the three tests. The test itself taps motor ability and does not overlap much with the other tests. These tests are highly recommended for a battery, although they correlate so highly with each other that unless theoretical issues (e.g., hemisphericity) are to be studied, using one is recommended.

The Reaction Time Tests. These tests exhibited stability but lower reliabilities for 1-Choice Reaction Time (.59). Only the 4-Choice Reaction Time is recommended as it does have higher stabilized reliability (.94), and covaries with the 1-Choice and 2-Choice Reaction Times.

Grammatical Reasoning. Because of technical difficulties, only the response latency was available for analysis, but for this score Grammatical Reasoning did show high reliability and fairly rapid stability. This test is recommended for a battery, particularly based on previous research (Kennedy, Wilkes, Lane, & Homick, 1985).

Associative Memory. This test required five sessions to stabilize and was reliable (r = .80). It correlates low with other tests possibly indicating its factor independence.

Pattern Comparison. This test was somewhat slow to stabilize, but exhibited high reliability in number correct as well as response latency. Therefore, this test is tentatively recommended and has performed well in previous studies (Kennedy, Wilkes, Lane, & Homick, 1985).

Air Combat Maneuvering. This test did not stabilize but we feel that it might have stabilized given more practice. Also, the test itself does seem to be a "motivating" task according to subjects' reports. Air Combat Maneuvering is recommended for further study.

Number Comparison. Number Comparison stabilizes within three sessions and exhibits acceptable reliability (0.92). Its correlation with other tests is moderate. This task is highly recommended.

Short-term Memory. This test stabilizes quickly and has high reliability for number correct and response latency. It is also highly recommended for use in a test battery.

Auditory and Visual Counting. These tests have their origins as vigilance tests (Jerison, 1955) and only provide 4-5 data points per minute in the complex versions and 1-2 per minute in the simple versions. Thus, the low reliabilities which were obtained in this study are not surprising. Additionally, while we find the Low and Medium Difficulty Counting tests to be unstable in this study, past research with longer administration times have shown them to be useful and stable measures with respect to vigilance and workload (Kennedy & Bittner, 1980). The tests also have the advantage of auditory or visual presentation. For these reasons we recommend the High Difficulty versions for further study.

In Table 5 may be found: 1) intertest correlations above the diagonal, 2) retest reliabilities (underlined) in the diagonal, and 3) below the diagonal intertest relationships corrected for attenuation due to unreliability based on the formula from Spearman (1904):

$$R = r_{12}/(r_{11}\ r_{22})^{1/2}$$

This formula allows one to estimate the amount of shared variance between two scores after correcting for their respective lack of reliabilities. Such a calculation provides a prediction of overlap versus prospective independence given perfectly reliable measures. From such an analysis inferences about factor richness may be made.

Several interesting relations are apparent in Table 5, particularly if one considers the corrected for attenuation correlations below the diagonal. First, is the fact that the auditory and visual counting tasks are interchangeable thus the choice of which to use depends solely on the conditions of the intended study. Second, the counting tasks share substantial variance with many other cognitive tasks particularly after correction, which implies that if their reliabilities were improved, perhaps by longer or repeated testing, either task would capture a substantial portion of total battery variance. Third, examining either the corrected or uncorrected intercorrelations between the tapping tasks and the reaction time tasks shows that tapping relates most to simple reaction time and less well as choice (cognitive complexity) is added. Therefore, in a simplified battery one should probably use tapping with four-choice Reaction Time, dropping simple Reaction Time because tapping is a simpler shorter task. Fourth, focusing on the Short-Term Memory task, one sees high corrected correlations with the Counting tasks, Four Choice Reaction Time, Number Comparison, Pattern Comparison, and Grammatical Reasoning; thus this task perhaps best represents or summarizes the higher cognitive functioning component of the battery. Many of the other intertask correlations, even after correction for attenuation were low, which, given the high reliabilities, implies that the tests of this menu tap different constructs or factors.

TABLE 5. INTERCORRELATIONS OF THE STABLE TESTS

	AUDHI	VISHI	RT1	RT2	RT4	STM	NCP	PCN	ASM	GRL	PTAP	THTP	NTAP
AUDHI	.78	.70	-.49	-.36	-.46	.63	.42	.50	.53	-.42	.47	.68	.44
VISHI	.90	.78	-.14	-.20	-.21	.68	.47	.34	.29	-.73	.50	.31	.51
RT1	-.72	-.21	.59	.68	.45	-.21	.12	-.26	.20	.23	-.60	-.46	-.39
RT2	-.42	-.23	.92	.94	.90	-.57	-.26	-.56	.21	.41	-.52	-.35	-.27
RT4	-.53	-.24	.45	.96	.94	-.68	-.51	-.72	-.06	.35	-.37	-.44	-.22
STM	.73	.79	-.28	-.60	-.72	.95	.80	.76	.38	-.71	.39	.36	.35
NCP	.49	.57	.16	-.28	-.55	.86	.92	.62	.44	-.56	.00	.08	-.08
PCN	.58	.37	-.35	-.60	-.77	.80	.67	.94	.45	-.52	.37	.54	.29
ASM	.67	.36	.29	.24	-.07	.44	.51	.52	.80	-.08	-.02	.33	-.01
GRL	-.48	-.84	.29	.43	.37	-.74	-.60	-.55	-.09	.96	-.39	-.28	-.23
PTAP	.53	.57	-.79	-.54	-.39	.40	.00	.38	-.02	-.40	.99	.53	.95
THTAP	.78	.36	-.61	-.37	-.46	.38	.09	.57	.38	-.29	.54	.97	.55
NTAP	.51	.59	-.51	-.28	-.23	.36	-.08	.30	-.01	-.24	.96	.56	.98

DISCUSSION

MENU OF TESTS

It is believed that too little attention is paid to evaluating tests prior to their use in studies of behavioral toxicology and occupational health. The 13 stable and reliable tests (scores) which we report in this study (Tables 5 and 6) are differentially stable and with generally high task definition. They comprise a cross-section of cognitive and psychomotor tasks, and because of the low relation of correlations between tasks and the very high reliabilities (average $\underline{r}$ = .89), a factor analysis in a large population is likely to reveal rich factor structures.

The findings of this study indicate that the core battery of five tests (Grammatical Reasoning, Pattern Comparison, and the Tapping series) are stable and reliable. Eight additional tests also were shown to possess the requisite metric properties. Because the tests are short (< three minutes) and easily administered, we would propose that the 13 tests can be customized variously to form batteries of differing lengths and composition to suit the individual investigator.

We would recommend that the five-test core battery could be easily augmented by Number Comparison, one of the Reaction Time (preferably 4-choice), and the two memory tests (Short-Term and Associative) for a 10-test (16-minute) battery. On the other hand, to conduct factor analysis studies, one might wish to select overlapping tests. Future studies should also examine the factor structure of such a battery in a larger population, perhaps with fewer replications.

For those tasks which showed slower stabilization times, it would probably be possible to double their practice time so that one hour could be allotted for baseline testing. It would therefore appear plausible to create a battery of tests of differing lengths and different numbers of tests for various purposes. To be most economical, one might start with tests showing the least overlap and add tests until the time available for testing is filled.

SELF-ADMINISTRATION

The field testing of this automated system indicates that the menu of tests can be successfully self-administered over repeated applications, outside a research laboratory environment. The research director need only initially instruct the subjects in the use of the battery, establish testing protocol and properly motivate the individuals involved in the study. We recognize that we cannot converge on whether the lack of stability of some of the tests and scores in this study is due to the self-administration or the tasks themselves or some interaction thereof. However, the present study produced 13 tests which met minimum requirements and this provides a useful nucleus. Additionally, based on the fact that those tests which were stable in previous studies (Kennedy, Wilkes, Lane, & Homick, 1985) were also stable here and with approximately the same metric features, we tentatively conclude the lack of stability in several remaining tests is likely a problem with the tests themselves rather than the self-administration methodology.

We think the notion of using self-administered portable microcomputer tests for fitness for duty has not yet been explored for persons in critical occupations (e.g., space, nuclear power plants), and in those cases where suspicion of a progressive disease (e.g., positive testing for human immunodeficiency virus) may occasion individuals to leave the workforce permanently and before performance changes have been shown to occur. The importance of opening data collection to laboratory free environments has broad applications.

TEST THEORY

The usual paradigm followed in studies of environmental stress and toxic agents entails exposure of one or more subjects to an intervention, then the individual's score under the treated and nontreated conditions is compared. However, implicit in such a design is that over and above the name of the test being the same, the behavioral element or construct being tapped must also be the same on each testing. It is well-known that learning a task may entail skills and abilities which are different from those required to perform the

task after it is well-practiced (Ackerman & Schneider, 1984) even to the extent, for example, that different structures in the brain appear necessary for these two functions. Horfl and Misantone (1976) showed that cutting temporal lobe connection interferes with learning and retention of tasks, but not with their performance per se. Therefore, a chief requirement for any test which is employed to reveal change due to treatment is that it be stable when no treatments are applied. Satisfaction of such a requirement permits "attribution of effect" when changes are found. Provocative evaluations of stability must be conducted not only for means and variance -- but for between session correlations, as well (Bittner et al., 1986; Jones, 1980). Only when a test demonstrates symmetry of the variance covariance matrix (Campbell & Stanley, 1963) is there assurance that neither the task nor the subject taking the test is changing (Alvares & Hulin, 1972). Very few attempts have been made to study these relations and, to our knowledge, no one else has made them a part of performance test battery development programs, although the requirement is well documented in the theory and practice of mental testing (Allen & Yen, 1979).

Another major criterion for test selection was that, if the test revealed individual differences, the retest reliability should be high (tests with no between-subject differences are acceptable, but virtually unknown). High reliability is desired because 1) low reliability suggests insensitivity, and 2) sensitivity experiments typically employ small numbers of repeatedly measured subjects. In this experiment a few tasks, which had previously been shown to have merit, either did not stabilize during the period of this experiment or possessed lower than desirable retest reliabilities. The reasons for this were not always due to the same causes and, we believe, in most cases the test could have qualified with longer administration periods.

Although in most cases the number correct and response latency scores are "purest" and preferred to percent correct, the latter serve as a check in determination of a subject's test taking strategy, which during repeated-measures testing with treatments, may change. For example, we have had experience (Kennedy, Dunlap, Banderet, Smith, & Houston, 1988) with a subject who rapidly pressed true/false response keys to generate a higher number correct score as an environmental stress influenced his ability to perform. Number Correct, in this case, increased and latency remained unchanged, but percent corrent went down to nearly 50%. It is highly recommended that in cases where subject motivation and test taking strategy are questionable, the percent score should be closely examined.

The literature which examines the interaction between human performance and the medium which is employed is not broad but the findings appear to be consistent. When correlational analyses have compared tests presented in computerized versus paper-and-pencil modes, the most usual finding is that the strongest correlations appear between the same tests in different media rather than among different tests in the same medium (Smith et al., 1983; Kennedy, Wilkes, Lane, & Homick, 1985; Kennedy, Dunlap, Wilkes, & Lane, 1985). When analogous questions have been raised regarding presentation of displayed information, as a function of the theoretically "appropriate" versus "less appropriate" channel (viz, vision, audition) the factor analytic findings (Wickens, Sandry, Vidulich, 1983) follow task structure rather than input pathway. This does not mean performance is the same with both media. The

number of bits of information which are handled can be improved by the channel selected for presentation, but the factorial representation of the basic human capacity to process the information appears to be largely unchanged by the medium selected. Stated differently, we believe that the data of this study show that the "message" of "medium" effect (McLuhan, 1966, p. 9) is weaker than that of the "factor" effect.

CONCLUSION

The data reveal that tests from what was previously the core battery (Grammatical Reasoning, Tapping, Pattern Comparison) correlate moderately with each other and resemble patterns of correlation from previous studies (Kennedy, Wilkes, Lane, & Homick, 1985; Kennedy, Dunlap, Jones, Lane, & Wilkes, 1985) when two to three factors were revealed in a small sample. Thirteen "new" tests which were used in this study included the counting family (six tests -- three each visual and auditory of varying difficulty). These latter tests were either unstable or not reliable enough for us to recommend them strongly. This was probably due to the low demand for responses, particularly in one- and two-channel monitoring. Although these tests appear different from other more traditional cognitive and information processing tasks, and have considerable face validity for monitoring watchkeeping tasks, their correlations after correction for attenuation imply considerable overlap with the constructs available in the other tests. However, the other several tests from this study which were stable and reliable can productively be used to form a middle-length battery. We would tentatively suggest one each of the Tapping, Reaction Time, Short-term Memory, Number Comparison, Pattern Comparison, Associative Memory, Grammatical Reasoning, and rexamination of the Counting series. Based on the results of this experiment, we would predict that with such a battery subjects, properly instructed can test themselves repeatedly and the tests will retain their good metric properties even over many repeated exposures.

REFERENCES

Ackerman, P. L., & Schneider, W. (1984, August). Individual differences in automatic and controlled information processing (Rep. No. HARL-ONR-8401). Champaign, IL: Human Attention Research Laboratory.

Allen, M. J., & Yen, W. M. (1979). Introduction to measurement theory. Monterey, CA: Brooks/Cole.

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

American Psychological Association. (1982). Ethical principles to the conduct of research with human participants. Washington, DC: Author.

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

Barrett, G. V., Alexander, R. A., Dovberspike, D., Cellar, D., & Thomas, J. (1982). The development and applications of a computerized information processing test battery. Applied Psychological Measurement, 6, 13-29.

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., 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.

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

Cronbach, L. J., & Furby, L. (1970). How we should measure "change" - or should we? Psychological Bulletin 74(1), 68-80.

Donders, F. C. (1968). On the speed of mental processes (W. G. Koster, Trans.). Acta Psychologica, 30, 412-431.

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

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 No. N00014-71-C-0117). Princeton, NJ: Educational Testing Service.

Essex Corporation. (1985). Automated performance test system. Orlando, FL: Brochure.

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

Hanninen, H., & Lindstrom, K. (1979). Behavioral test battery for toxico-psychological studies used at the Institute of Occupational Health in Helsinki (2nd ed.). Helsinki, Finland: Institute of Occupational Health.

Horfl, J. A., & Misantone, L. J. (1976). Visual discrimination impaired by cutting temporal lobe connections. Science, 193, 336-338.

Jeanneret, P. R. (1988, February). Position requirements for space station personnel and linkages to portable microcomputer performance assessment. Orlando, FL: Essex.

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.

Jones, M. B. (1970a). A two-process theory of individual differences in motor learning. Psychological Review, 77(4), 353-360.

Jones, M. B. (1970b). Rate and terminal process in skill acquisition. American Journal of Psychology, 83(2), 222-236.

Jones, M. B. (1980). Stabilization and task definition in a performance test battery (Final Rep., Contract N00203-79-M-5089). New Orleans, LA: U.S. Naval Aerospace Medical Research Laboratory. (AD A099987)

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

Kennedy, R. S., Dunlap, W. P., Bandaret, L. E., Smith, M. G., & Houston, C. E. (1988, in press). Cognitive performance deficits occasioned by a simulated climb of Mount Everest: Operation Everest II. Aviation, Space, and Environmental Medicine.

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., Dunlap, W. P., Wilkes, R. L., & Lane, N. E. (1985, October). Development of a protable computerized performance test system. Paper presented at the 27th Annual Meeting of the Military Testing Association. San Diego, CA.

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.

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

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

McLuhan, M. (1966). Understanding media: The extensions of man. New York: McGraw-Hill Paperback Edition.

Michael, J. M. (1982). The second revolution in health: Health promotion and its environmental base. American Psychologist, 37, 936-941.

NEC Home Electronics (USA), Inc. (1983). NEC PC-8201A users guide. Tokyo, Japan: Nippon Electric Co. Ltd.

Nicholi, A. M., Jr. (1983). The nontherapeutic use of psychoactive drugs: A modern epidemic. The New England Journal of Medicine, 308(16), 925-933.

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

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.

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

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.

Turner, C. E. (1983). Cannabis: The plant, its drugs, and their effects. Aviation, Space, and Environmental Medicine, 54, 363-368.

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

Weiss, B. (1983). Behavioral toxicology and environmental health science: Opportunity and challenge for psychology. American Psychologist, 38(11), 1174-1187.

Wickens, C. D., Sandry, D., & Vidulich, M. (1983). Compatibility and resource competition between modalities of input, central processing, and output: Testing a model of complex task performance. Human Factors, 25, 227-248.