DOT/FAA/AM-92/11

A Candidate Automated Test Battery for Neuropsychological Screening of Airmen: Design and Preliminary Validation

Office of Aviation Medicine
Washington, D.C. 20591

Robert D. O'Donnell
Jerry R. Hordinsky
Sudahar Madakasira
Samuel Moise
Debra Warner

Civil Aeromedical Institute
Federal Aviation Administration
Oklahoma City, Oklahoma 73125

February 1992

Final Report

U.S. Department
of Transportation
Federal Aviation Administration

NOTICE

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/AM-92/11		

4. Title and Subtitle	5. Report Date
A CANDIDATE AUTOMATED TEST BATTERY FOR NEUROPSYCHOLOGICAL SCREENING OF AIRMEN: DESIGN AND PRELIMINARY VALIDATION	February 1992
	6. Performing Organization Code
	8. Performing Organization Report No.

7. Author(s) R. D. O'Donnell, Ph.D., J. R. Hordinsky, M.D., S. Madakasira, M.D., S. Moise, Ph.D., D. Warner, M.S.

9. Performing Organization Name and Address
FAA Civil Aeromedical Institute
P. O. Box 25082
Oklahoma City, OK 73125

10. Work Unit No. (TRAIS)	11. Contract or Grant No.	13. Type of Report and Period Covered

12. Sponsoring Agency Name and Address
Office of Aviation Medicine
Federal Aviation Administration
800 Independence Avenue, S.W.
Washington, D.C. 25091

14 Sponsoring Agency Code

15. Supplementary Notes
Work was accomplished under contract number DTFA-02-87-C-87070

16. Abstract

A panel of the American Medical Association convened by the Federal Aviation Administration recommended that a computerized test of cognitive function be developed that would detect significant cognitive impairments that might otherwise go unrecognized during a routine physical examination. In response to this need, a computerized test battery, based on current cognitive theory, has been developed that provides a brief screening for disturbances in higher-level cognitive function. This battery is not designed to replace the traditional observational methods used by the physician, but rather to enhance diagnostic sensitivity in areas not currently well covered. The battery operates in a "step" fashion, providing a generalized, non-specific screen at the first level, with two increasingly more specific screens if that level is failed. The output of the battery is a verbal protocol to the examiner presenting a series of "rule out" recommendations for further diagnostic testing. In this report, the background and composition of this test are described, and the results of three initial validation and sensitivity studies are reported. The present test development suggests the utility of transferring previously expensive and more complex diagnostic approaches to a computerized testing and decision process. This development of an automated approach to cognitive function testing was one of three sponsored by the FAA Office of Aviation Medicine during 1987 - 1990.

17. Key Words	18. Distribution Statement
Neurological screening, Computerized test, Physical examination, Psychiatric screening, Mental status examination	Document is available to the public through the National Technical Information Service Springfield, Virginia 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	18	

ACKNOWLEDGEMENT

The authors wish to express their appreciation to the staff of the Veterans Administration Hospital, Dayton, Ohio, for their cooperation in identifying and scheduling subjects for these studies. In particular, Dr. Thomas Matthews, Ms. Trudy Cortez, Dr. Dennis Johnson, and Dr. Vincente Callejo took time from their already overcrowded schedules to assist in these efforts.

We also wish to acknowledge the creative and administrative efforts of Ms. Karen Peio and Mr. Mark Crabtree in early phases of experimentation, and Mr. R. L. O'Donnell for his administrative and data analysis efforts.

Finally, we express our appreciation to all of the subjects, whose unselfish cooperation made these studies possible.

AUTHORS' AFFILIATIONS

Robert D. O'Donnell, Ph.D.
Samuel Moise, Ph.D.
Debra Warner, M.S.

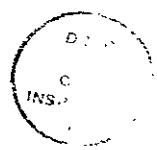
NTI, Incorporated
4130 Linden Avenue, Suite 235
Dayton, Ohio 45432

Jerry R. Hordinsky, M.D.

Civil Aeromedical Institute
Federal Aviation Administration
Oklahoma City, Oklahoma 73125

Sudahar Madakasira, M.D.

Department of Psychiatry and Human Behavior
University of Mississippi Medical Center
2500 North State Street
Jackson, Mississippi 39216



CONTENTS

A CANDIDATE AUTOMATED TEST BATTERY FOR NEUROPSYCHOLOGICAL SCREENING OF AIRMEN: DESIGN AND PRELIMINARY VALIDATION

INTRODUCTION

The neurological screening tests carried out routinely in the course of an airman physical certification examination are designed to detect a broad range of sensory-motor abnormalities, with particular emphasis on the cranial nerves. This examination may or may not be accompanied by a relatively informal mental status examination exploring psychiatric and cognitive functions (Siassi, 1984). However, with increasing concerns about the need to assess higher mental functions, it is recognized that the scope and sensitivity of neuropsychological aspects of the examination must be expanded. For instance, a panel of the American Medical Association recently convened by the Federal Aviation Administration recommended that a computerized test of cognitive function be developed "... that would detect significant cognitive impairments that may otherwise go unrecognized during a routine physical examination." (AMA, 1984).

The problem is that current neuropsychological screening and mental status examinations were designed to detect symptoms of relatively severe sensory, motor, or cognitive pathology. While the tests appear relatively good in detecting such severe organic illness, ranging from 60 to 70 per cent accuracy (Webster, Scott, Nunn, McNeer, and Varnell, 1984), they were not intended to be so sensitive that they could be used as early indicators of disturbances of higher-level mental function. The "cognitive function" portion of the traditional mental status examination is typically limited to observing the patient's orientation for time and place, knowledge of birthdate and age, and some historical or geographical reference, such as the name of the current President or the location of the test (Siassi, 1984). Although this may be supplemented by observing the patient's form of thinking (logical or iliogical), and ability to abstract, it is easy to see that this examination does not challenge higher mental functions. In fact, it has undergone little change from the time it was introduced by Adolf Meyer in 1902, despite significant theoretical advances in the field of cognitive psychology (Gardner, 1987).

In response to this need, a computerized test battery based on current cognitive theory, has been developed that provides a brief screening for disturbances in higher-level cognitive function. This report describes the background and composition of this test, and the results of initial validation and sensitivity studies.

Background

Many attempts to make the traditional mental status examination more objective have been carried out. A good sampling of these has been described by Nelson, Fogel, and Faust (1986). One, the Mini-Mental Status Examination (MMSE) (Folstein, Folstein, and McHugh, 1975) is of particular interest since the AMA Committee referred to above, after considering existing test procedures, recommended to the FAA that the MMSE be used in the routine cognitive screening of candidate airmen until a more definitive test battery could be developed (AMA, 1984). The MMSE is a general-purpose cognitive screening test consisting of 11 items and requiring 5 to 10 minutes to administer. The tests measure orientation to time and place, registration (naming 3 objects and remembering them), attention and calculation (serial-seven subtraction), recall (remembering the 3 objects repeated above), language (naming, repeating, and following commands), motor behavior (eye closing), sentence production, and copy design. The patient's level of consciousness is also subjectively evaluated along a continuum from alert to coma.

Nelson, et al., (1986) report the results of 35 publications dealing with the MMSE. Five tests of reliability were reported revealing a range from .83 to .99 in psychiatric, neurological and mixed patients. A total of 11 validation studies covering 859 subjects has been carried out, with the percentage of correct classification ranging widely, depending on the pathology involved. For instance, non-psychotic psychiatric inpatients without diagnosed organic mental disorders, and patients with focal right hemisphere lesions almost all pass this test, while patients suffering from depression with cognitive impairment usually fail. Anthony, LeResche, Niaz, von Korff, and Folstein (1982) report an overall false positive rate of 39 percent and a false negative rate of 5 percent. The correlation between the MMSE and the Wechsler Adult Intelligence Scale is reasonably high (between .55 and .78 for the verbal portion, and between .55 and .66 for the performance portion) (Dick, Guiloff, Stewart, Blackstock, Bielawska, Paul, and Marsden, 1984). Subsequent to the AMA recommendation, the MMSE was validated in three studies with respect to its ability to discriminate between civil pilots and neurological or psychiatric patients. In all three, results were extremely disappointing, with false negative rates as high as 96 percent (LeRoux, 1988).

Foundations of the new test battery.

The test battery proposed here evolved from a relatively new theoretical orientation, and utilizes a "step" procedure to minimize testing time. These approaches are described briefly below.

While traditional neuropsychological tests have utilized theoretical clinical or empirical approaches to test construction, the present battery was developed with a specific theoretical position in mind. Most recent formulations of the nature of human cognition postulate that it is multi-dimensional, i.e., separate processing mechanisms exist for general categories of cognitive function. This multi-processor hypothesis proposes that two activities can be conducted without mutual impairment, as long as each one utilizes a different information processing structure (Allport, 1980: see Colley and Beech, 1989 for a review). Based on this kind of data, Wickens (1984) proposed a "multiple-resources" theory, suggesting that there are at least three kinds of resources, each varying in two dimensions. The first resource involves sense modality, and primarily involves visual versus auditory processing. The second resource divides the above into early and late stages of processing. The third resource addresses the type of central processing carried out (spatial or verbal).

In the present test battery development, the multiple resources theory provided a basic starting point. The goal was to use specifically targeted tests to sample as many of the resources postulated by Wickens as possible. To this end, spatial and verbal functions requiring various memory demands and processing sequences were included. In addition, tests of psychomotor control and some of the best traditional clinical tests were included to provide the broad-based screening desired for the first tests in the battery.

The second relatively new characteristic of the present test battery was incorporated to answer the need for a brief screening test which was also of some diagnostic use to the clinician. The "step" approach, as described by Russell (1984), was adopted for this purpose. In this, the battery is organized into steps. If the person fails tests in the first step of the battery, the next set of tests is administered to verify and elaborate the indications of the first step. Thus, each step acts as a screening battery for the next, more detailed step.

Versions of such a step approach to testing appear to be gaining in popularity. Tarter and Edwards (1986), for example, suggest that brief screening tests be used to explore "core elements of a neuropsychological examination" (attention, memory, perception, language, visual-spatial, and psychomotor processes). If indicated by results on these tests, a second stage would give a standard "subbattery" to specific individuals. Problems at this second level would signal the need for highly individualized testing. This approach has been incorporated into the 1.5 hour Pittsburgh Initial Neuropsychological Test System (PINTS) (Goldstein, Tarter, Shelly, and Hedgedus, 1983). However, as used to date, the step approach has not been truly incorporated into a computer-driven brief neuropsychological screening battery of the type required in the present effort.

Brief description of the Neuropsychological Test Battery (NTB).

Recognizing that the development of a test battery is an iterative process, a preliminary set of candidate tests was selected. These tests were implemented in a "breadboard" fashion, using paper-and-pencil tests and a Commodore computer. The first validation experiment utilized this version (LeRoux,1988). Based on the results of this experiment, a revised first candidate version (1.0) was created, consisting of a different combination of the original tests. This version was administered to 121 subjects and, based on these analyses, a second-generation breadboard version (1.1) and a fully-computerized version (2.0) were developed and subjected to preliminary experimental validation. In this section, the tests comprising these various versions are described.

The original list of candidate tests considered in the preliminary version consisted of the following:

1. **Trail Making Test.** This is a test of "visual-conceptual and visuomotor tracking" (Lezak, 1983). In the first part of the test (Trails A) 25 numbered circles are to be joined in sequence. In the second part (Trails B) the 25 circles are numbered 1 to 13 and A to L, and they are to be joined in an alternating sequence. The test has consistently proven to be one of the best general screening instruments for diffuse brain injury (Spreen and Benton, 1965). In addition, it has been shown to be decremented in chronic alcoholics, in certain neurological conditions, and in psychiatric conditions (Lezak, 1983).

2. **Symbol Digit Substitution Test.** This modification of the digit-symbol subtest from the Wechsler *intelligence scales is based on the work of Smith* (1968). It requires the subject to substitute numbers for geometric symbols. It appears to require visual perceptual, visual scanning, and attention allocation resources. It is reported to be more consistently sensitive to brain damage than any other Wechsler Adult Intelligence Scale subtest, and to show decrement even when damage is minimal.

3. Color-Word Test. This test, modified from the original Stroop Test (Stroop. 1935) requires the subject to name the color in which a word is written, even though the word may be the name of that color, or of a different color. It is a measure of the speed with which a person can inhibit an overlearned perceptual set (reading the word) and conform to changing demands. As such, it appears to tap several of the central processing and response organization resources of the multiple-resources model.

4. **Unstable Tracking Test.** In this test, the subject must keep a computer-generated "target" centered with a tracking knob (version 1.1) or a joystick (version 2.0), while the computer generates offsets for the target. This test has considerable content validity as a sensitive visual-motor coordination test. In pretests of the current battery (Leroux, 1988) this test proved to be one of the best general discriminators between normal and pathological groups.

5. **Continuous Performance Test.** This Dynamic Memory Test is modeled after procedures described by Moore and Ross (1963) and Hunter (1975). The basic design for the present version of the test was developed by Shingledecker (1984) as part of the Criterion Task Set for the U. S. Air Force. The test requires the subject to note the bottom number of a fraction. When a new fraction appears, the subject must respond by saying whether the top number is the same as the previous bottom number. However, the new bottom number must first be noted, because as soon as a response is given, the original fraction is replaced by a new one. Again, elements of numerical central processing and response inhibition are probed by this procedure.

6. **Verbal Thinking Test.** This test is based on the paradigm developed by Posner (1978), and involves having the subject classify two letters of the alphabet by each of two rules. One rule involves physical identity alone (whether both are the same letter in the same case). The other involves a semantic rule (whether both are vowels or consonants). The test places high demands on semantic memory, and on rule-based behavior.

7. **Arithmetic Test.** This is a simple test of ability to carry out several addition and subtraction functions rapidly. It has been adapted from the Unified Tri-Services Cognitive Performance Assessment Battery (UTCPAB) (Perez, Masline, Ramsey, and Urban, 1987), and appears to probe specific numerical, logical, and attention allocation functions.

8. **Interval Production Test.** This test is based on the work of Michon (1966), and requires the subject to tap at a regular rate of two to three per second for three minutes. Interest is in the variability of the tapping. It appears that the test may measure psychomotor stability (possibly involving the reticular-cerebellar axis), and should be sensitive to disruptions due to either organic or functional problems.

9. **Spatial Thinking Test.** This test dates from an original concept described by Fitts (1956) and is modified by Shingledecker (1984). A four-bar histogram is presented. After 3 seconds it is removed and replaced (after a delay) with another histogram rotated either 90 or 270 degrees. The subject must decided whether the second histogram is the same as the first. Intact spatial memory is required, as well as ability to mentally manipulate spatial symbols.

10. **Short-term Memory/Retrieval Test.** The paradigm proposed by Sternberg (1969) is used to probe short-term memory retrieval processes (including sensory/perceptual and motor functions). This test involves determining whether a "probe" letter of the alphabet is a member of a previously memorized target set. Short-term retrieval processes are required by this test.

11. **Visual Monitoring.** This test requires the subject to monitor four dials (similar to aircraft dials) to detect a randomly-occurring bias in one of them.

12. **Logical Reasoning Test.** This is a version of the logical reasoning test proposed by Baddeley (Baddeley and Liberman, 1980). A series of symbols are presented, along with a verbal description of the logical relationships between them. The subject must determine whether the logical relations described are true or not with respect to the presented symbols. The test assesses a broad range of higher level cognitive functions.

13. **Zung Self-Rating Depression Scale** (Zung, 1965). This is a 20-item scale which yields an overall depression index, as well as sub-scores on affect, physiological disturbance, and psychomotor disturbance.

14. **Manifest Anxiety Scale.** This is a 28 item self-report scale designed to detect symptoms of "anxiousness," primarily as manifested in autonomic activity.

15. **Shipley-Hartford Retreat Scale.** This two-part test consists of a vocabulary section and an abstract reasoning section. The vocabulary test was used in early testing only to establish that the subject is functioning at an acceptable intellectual level.

Software.

Versions 1.0 and 1.1 (the breadboard versions) were implemented on Commodore computers. Version 2.0 was created in QuickBasic, and is compatible with IBM XT or higher computers. QuickBasic is a common language which is familiar to most computer users. Thus, it is easily modifiable. The program automatically presents tests, evaluates subject performance at each level, decides whether to present subsequent levels of tests, and prints to the screen a code that tells the examiner the results of the examination. All results, of course, are saved. Details of the computer program are described in more detail in Moise, O'Donnell and Hordinsky (in preparation).

Hardware.

The battery is configured to run on an IBM XT, AT, or true clone with 512K memory, one 360K floppy disk, either an EGA graphics card or a Sigma Designs Color 400+ graphics card, and a color monitor that supports the selected color graphics card. This configuration is a reasonably "standard" PC system of the type which exists in many physicians' offices.

Preliminary validation studies.

The long and demanding process of criterion and predictive validation will clearly take several years to complete. However, as an initial attempt, three validation studies involving 242 subjects have been carried out to provide the initial assessment of the proposed battery, as well as to provide a model for subsequent validation studies. A preliminary study used 121 subjects to assess the Mini-Mental Status Exam of Folstein, Folstein, and McHugh (1975), and to provide basic data on the preliminary set of performance tests. This study, reported in LeRoux, (1988), provided the experimental basis for selection of the initial tests in version 1.0 of the NTB. The next two studies provided important clues with regard to subsequent modifications of the battery, and are described in the present report.

Materials and Methods.

Subjects. A total of 121 subjects were tested in the first study reported here, with 81 individuals in the "non-pathology" group (no history of psychiatric or neurological pathology) and 40 subjects in the "pathology" groups. In the non-pathology group, 41 of the subjects were active pilots, and 40 were non-pilots fulfilling the group criteria. Twenty of the subjects in the overall cohort of the first study were used again in a second study, as explained later.

The non-pathology subjects all agreed to participate without compensation. The pathology subjects were recruited through the local Veterans Administration Hospital (VAH) Center. These VA subjects were paid at the rate of $5.00 per hour of participation. No attempt was made to control or interfere with the normal medication for any subject. Patients were, for the most part, well-controlled on their present medication. They thus represented a clinical population currently displaying only marginal symptoms.

The pathological groups, and the number of subjects in each sub-category finally included in the test sample for the first study, are described below:

1. **Substance abuse.** This group included 22 subjects currently being treated for alcoholism, and/or drug dependency. All of these subjects were more than 90 days post-detoxification, by clinical record.

2. **Seizures.** Included in this group were eight individuals currently being treated in a hospital neurology department, all of whom had been diagnosed as having seizure disorders from various causes.

3. **Depressives.** Included in this group were hospital inpatients (2) and outpatients (8) who carried a primary diagnosis involving depression, and who were currently being treated by psychotherapy and/or medication for that condition.

For the second experiment described below, 20 subjects who had been evaluated with versions 1.0 and 1.1 were retested with the version 2.0. This included 5 subjects from the pilot group and 5 from the non-pilot-normal group, in addition to 10 subjects from the pathological groups. These latter subjects consisted of 4 from the substance abuse category, 3 from the neurological category, and 3 from the depression category.

Procedures for first study.

Based on the results of the preliminary study, candidate tests for each of three levels of the battery were selected, and this was designated as version 1.0 of the NTB. The tests selected are shown in Table I.

All subjects were given all of the Level 1 tests. After they were finished, their scores were inspected and compared to pre-established "pass-fail" criteria on each test. These preliminary criteria were deliberately set to be

TABLE I. VERSION 1.0 OF THE NTB

LEVEL 1	LEVEL 2	LEVEL 3
TRAILS A	CONT. PERFORMANCE	SPATIAL THINKING
TRAILS B	VERBAL THINKING	MEMORY TEST
SYMBOL DIGIT	ARITHMETIC	VISUAL MONITOR
COLOR WORD	INTERVAL PRODUCTION	LOGICAL REASON.
UNSTABLE TRACKING		ZUNG DEPRESSION
		MANIFEST ANXIETY
		SHIPLEY SCALE

harder to "pass" than was expected for the final criterion measures. In this way, it was assured that all subjects who would eventually be failed by the final (less rigorous) criteria would also fail in this preliminary screening. In any case, if the subject failed any test at Level 1, all of the tests at Level 2 were administered. The same logic as above was used to establish Level 2 "pass-fail" criteria, and if the subject failed any test at Level 2, all Level 3 tests were administered.

All computer-generated tests in versions 1.0 and 1.1 were presented on a Commodore SX-64 computer, using a 12-inch Commodore color monitor. Subjects received immediate feedback after each test. In addition to the computer-generated tests, several commercial paper-and-pencil tests were administered in these versions. These were the Trails Test (forms A and B), the symbol-digit test, and the Shipley Scale. These tests were given in their standard commercial forms, using the directions and norms provided by the test authors.

Procedures for the second study.

From the entire group of 121 subjects who had participated in the first experiment, 45 randomly selected subjects were contacted by the experimenter, and requested to participate again. The first 20 to accept in the appropriate categories were used as the subjects. Except for the testing sequence and the completely computerized administration, all procedures were identical to the original test administration. As in the first study, subjects in the pathological groups were paid for participation, while "normal" subjects received no compensation. Every attempt was made to maintain the same motivation level as in the first study, and it was felt that conditions between the two test administrations were as identical as possible.

RESULTS AND DISCUSSION

Results of the first study.

A total of 62 individual measures (e.g., reaction times, percent correct scores, and standard deviations) were generated by the candidate battery. Summary data for each subject were analyzed in several ways. One-way analyses of variance (ANOVA-independent groups with unequal N) were performed on each of the dependent variables, based on group membership in any of the experimental groups. It is recognized that, with the large number of analyses thus carried out, a given alpha level is not protected, and therefore individual significances revealed in these analyses may not be precise, although a protection factor was used. Meaningful trends should, however, be revealed.

Age. There were proportionately more older subjects represented in the pathological groups, especially in the age ranges over 45 years. Mean age for the pathological groups was 48.15 years, as compared to 40.27 years for the non-pathological groups ($p<.001$). Closer inspection of the data, however, indicated that the age factor might not be as important as it first appeared. The major difference in age among the groups was between the non-pilot normal group and all others. In fact, the non-pilot group's average age was 34.8 years, whereas each of the other groups averaged between 44.3 (neurologicals) and 53.8 years (depressives), with a mean of 47.09 years. This compared to a mean age of 45.6 years for the pilot group. Thus, the pathological groups were not different from each other, or from the pilot subjects. Nevertheless, age was included as a variable in all subsequent analyses reported here. In addition, an analysis of covariance was performed between pathological and non-pathological groups on all of the dependent variables reported later, using age as the covariate. In no case was the basic statistical significance of any result changed (although, of course, significance levels were reduced somewhat).

TABLE II. NTB VARIABLES AND THEIR DISCRIMINATION LEVELS FOR THE EXPERIMENTAL GROUPS

TEST	VARIABLE	p-VALUE
SYMBOL DIGIT	SCORE	<.0001
	% CORR	.332
TRAILS A	TIME	<.0001
TRAILS B	TIME	<.0001
STERNBERG MEMORY RETRIEVAL		
	RT - SET 1	.02
	RT - SET 2	.03
	RT - SET 4	.009
	SD - SET 1	.02
	SD - SET 2	.036
	SD - SET 4	.004
	% CORR - SET 1	.091
	% CORR - SET 2	.202
	% CORR - SET 4	.765
	RT - TOTAL	.020
	% CORR - TOTAL	.227
	SLOPE	.038
	INTERCEPT	.029
DYNAMIC MEMORY (CONT. PERFORMANCE)		
	RT	.002
	SD	.013
	% CORRECT	.005
VERBAL THINKING TEST		
	RT - PHYSICAL	.009
	SD - PHYSICAL	.007
	% CORR - PHYSICAL	.0001
	RT - CATEGORY	.005
	SD - CATEGORY	.016
	% CORR - CATEGORY	.01
	PHYS. - CAT. DIFFERENCE	.036
	TOTAL % CORRECT	.001
LOGICAL REASONING		
	RT	.04
	SD	.060
	% CORRECT	.009
SPATIAL PROCESSING		
	RT	.213
	SD	.749
	% CORR	.108
STROOP COLOR WORD		
	RT - CONFLICT	.0003
	SD - CONFLICT	.0006
	% CORR - CONFLICT	.244
	RT - NON CONFLICT	<.0001
	SD - NON CONFLICT	.0003
	% CORR - NON CONFLICT	.536
	CON-NON CONFLICT RT	.083
	MEAN % CORR	.082

TABLE II (Continued). NTB VARIABLES AND THEIR DISCRIMINATION LEVELS FOR THE EXPERIMENTAL GROUPS

TEST	VARIABLE	p-VALUE
VISUAL MONITORING		
	RT	.055
	SD	.051
	HITS AFTER TIMEOUTS	.365
	FALSE ALARMS	.937
	TOTAL HITS	.003
	MISSES	.003
	% CORRECT	.003
INTERVAL PRODUCTION TEST		
	DURATION	.05
	SD	.007
	IPT	.087
UNSTABLE TRACKING		
	ERROR SCORE	<.0001
	EDGE VIOLATIONS	<.0001
ARITHMETIC		
	RT	.03
	SD	.103
	% CORR	.609
	NUMBER ATTEMPTED	.006
	NUMBER CORRECT	.003
ZUNG DEPRESSION SCALE		
	SCORE	<.0001
MANIFEST ANXIETY SCALE		
	SCORE	<.0001
SHIPLEY SCALE		
	SCORE	.021

Therefore, although age must be considered as a moderator in any future analysis, it does not appear to be the major determinant of the results to be presented below.

Sex. Similarly, there were very few female subjects available in the selected populations of civil pilots, and none for the VA patients. However, it was possible to look at sex differences in performance within the two "normal" groups. Analyses of variance were performed on all variables between the eight female subjects and the 73 male subjects. These revealed only 3 of the 62 variables significant at an alpha level of .02 or below (the .05 protected alpha level). It is therefore unlikely that there are true sex differences in performance on any of the tests.

"Intelligence." The Shipley score provides a crude measure of intelligence, and these were significantly different among the groups at the .02 alpha level. The pilot group scored higher than all other groups. The depressives (29.5) and the substance abuse subjects (28.7) were not different from each other, but were different from the neurologicals (26.7). In effect, these results suggest again that caution must be exercised in interpreting differences among experimental groups. In the final analysis, many of the measures to be used in any test battery will probably have to be moderated with an age and intelligence correction factor.

Test variables. The first analysis involved performing independent ANOVAs for the five experimental groups (pilots, non-pilot normals, depressives, substance abuse, and neurologicals) on all 62 of the dependent variables, plus the age variable. Results of these analyses are presented in Table II, and reveal that 42 of the 62 variables (68%) showed F-ratios with probability values less than .05. This number of significant results clearly suggests that the combination of tests in the battery will be able to differentiate among the experimental groups to a considerable degree.

TABLE III. VERSION 1.1 OF THE NTB

LEVEL 1	LEVEL 2	LEVEL 3
TRAILS A	LOGICAL REAS. (% COR.)	MEMORY (SLOPE)
TRAILS B	DYNAMIC MEMORY (S.D.)	ZUNG DEPRESSION
SYMBOL DIGIT	ARITHMETIC (ATTEMPTS)	MANIFEST ANXIETY
TRACKING		DYN. MEMORY (R.T.)

A goal of the first validation study was to arrive at a second-level battery of tests based on the results obtained among the pathological and non-pathological groups. Therefore, once the sensitivity of each test was established, the next step was to explore the nature of these differences and to select the specific tests and variables which would give the best diagnosticity. As a start, post-hoc (Newman-Keuls) tests of all significant variables were carried out for each of the proposed levels of the battery. The results of these analyses were then inspected to arrive at a preliminary list of variables which appeared to yield optimum differentiation among the experimental groups. These tests were then aggregated into a revised battery (designated version 1.1) to produce an optimal classification of subjects based on this sample. Optimization is appropriate at this early stage of test development, rather than employing a split-half or jack-knife procedure to cross-validate the tests selected. In view of this, it is obviously inappropriate to overinterpret sophisticated statistical analyses.

The above comparisons among the individual experimental groups revealed that, as expected, the Level 1 tests were generally excellent at differentiating between pathological and non-pathological groups, but were not very discriminating among the pathological groups. This is appropriate for a first-level screening procedure. Further inspection revealed that one of the first level tests (the Stroop Color Word Test) was not contributing as much to this differentiation as the other three Level 1 tests. For this reason, it was decided to eliminate the Stroop Test from the battery.

Similar analyses of each of the Level 2 and Level 3 tests originally proposed resulted in several other changes. The interval production test failed to identify normals or any pathology group. On the other hand, both the continuous memory test and the verbal thinking test appeared to be more discriminating among pathology groups than was originally hypothesized. Thus, they both appeared more appropriate for Level 3 than for Level 2. In their place, the logical thinking and one variable from the memory test (the standard deviation) appeared to give the best second-level differentiation between pathologicals and non-pathologicals, and these tests were therefore moved into Level 2.

In summary, the revised version (1.1) of the battery included most of the tests from the originally proposed battery, but made several changes in the order of test administration. The tests and variables included in this version of the battery are shown in Table III.

Having maximized the tests and variables which appear to have the ability to differentiate pathology, it was next necessary to develop the cut-off scores and decision logic to be used in automating the screening process. The test set data were used to produce a set of candidate cut-off scores using multiple criteria. A code was then developed that capitalized on the differing diagnostic levels of the battery. This code, along with the scoring paths used to generate it, is shown in Table IV. It is recognized, of course, that these statements will be modified as a result of cross-validation studies and further experience with the battery. Whatever their final form, these will inform the clinician of the level of the subject's performance, and the probable diagnostic implications of each performance level. Appropriately, the statements allow the clinician a considerable degree of latitude in determining the final disposition. They are, however, tied rigorously to the experimental results.

Classification accuracy.

Having created the classification algorithm based on the data from the present experiment (the training set), one would expect that the classification accuracy of this algorithm will be optimal and, hopefully, quite high. This proved to be the case in the present experiment. Version 1.1 of the NTB successfully identified 95 percent of the true positives (5 percent false negative rate). Further, the test battery may do equally well in eliminating the excessive cost associated with a high false positive rate. At Level 1, only 14 (17%) out of 81 subjects were "incorrectly" passed on to Level 2 testing. Of these, only 7 (8.6%) failed the Level 2 tests and were passed on to Level 3. Of these seven subjects, one passed all the tests, resulting in an overall false positive rate of 7.4%. These rates (5% false negatives and 7% false positives) are, of

TABLE IV. RECOMMENDED DIAGNOSTIC MESSAGES FOR VARIOUS LEVELS OF PERFORMANCE ON THE NTB

1. IF SUBJECT PASSES ALL TESTS IN LEVEL 1.	This subject has demonstrated performance on all tests in the screening battery that is within the limits of subjects not diagnosed as having neurological insult, affective disorders, or chronic substance abuse problems.
2. IF SUBJECT FAILS ONE TRACKING TASK AND NO OTHER TEST.	This subject has passed all tests in the screening battery except a demanding test of visual-motor coordination. Many normal subjects fail this test. Therefore, if the subject shows no clinical signs of visual-motor abnormalities, the screening battery should be considered to have been passed.
3. IF THE SUBJECT FAILS ONE TRACKING TASK AND ANY ONE OF THE OTHER THREE TESTS IN LEVEL 1.	This subject shows an overall pattern that is consistent with subjects not having diagnosed neurological insult, affective disorders, or chronic substance abuse problems. However, at least two of the individual tests were failed. While this failure rate is not diagnostic, it is recommended that increased attention be given to clinical signs of psychiatric or neurological abnormality in subsequent examination. If no such signs are present, the subject should be passed.
4. IF THE SUBJECT IS PASSED ON TO LEVEL 2, BUT PASSES ALL TESTS AT THAT LEVEL.	This subject shows an overall pattern that is consistent with subjects not having diagnosed neurological insult, affective disorders, or chronic substance abuse problems. However, the subject has shown a performance pattern that is weak in one or more skills. Such weaknesses have not usually been associated with psychiatric or neurological problems. However, increased attention should be given to clinical signs of such problems in subsequent examination. In the absence of such signs, the subject should be passed.
5. IF THE SUBJECT IS PASSED ON TO LEVEL 2 AND FAILS ANY TESTS AT THAT LEVEL, BUT THEN PASSES ALL TESTS AT LEVEL 3.	This subject should be screened carefully for neurological or psychiatric problems. The test battery suggests that the individual has a performance or skill deficit that is shared by many individuals with such problems. However, there is no specific indication of such problems in the responses of the subject. Therefore, if clinical examination is totally negative, the individual should be passed - otherwise, the subject should be referred.
6. IF THE SUBJECT IS PASSED TO LEVEL 3, AND FAILS ONE OR MORE TESTS AT THAT LEVEL (EXCEPTING THE SPECIFIC CASES NOTED BELOW).	This subject shows a pattern of performance that has been demonstrated by individuals diagnosed as having psychiatric, neurological, or substance abuse problems. Further testing is therefore strongly indicated. It is recommended that this individual be given a more intensive neurological and psychiatric screening and, if indicated, that further referral for specialized testing be made.
7. IF THE SUBJECT IS PASSED ON TO LEVEL 2, AND THEN FAILS: DECISION AND MEMORY SD SET 2, AND DECISION AND MEMORY SLOPE IN LEVEL 3.	This subject shows a pattern of performance that has been seen in several individuals diagnosed as having depressive disorders. While there are many other possible explanations for this pattern, it is recommended that increased screening for psychiatric disturbance should be carried out on this individual.
8. IF THE SUBJECT IS PASSED ON TO LEVEL 2, AND THEN FAILS: DYNAMIC MEMORY ALONE (NO OTHER LEVEL 3)	This subject shows a pattern of performance that has been seen in some individuals diagnosed as having substance abuse problems. There are many other possible explanations for this pattern, and the data on this relationship are tentative. Therefore, while it is recommended that increased screening for substance abuse should be carried out on this individual, a negative clinical finding should be considered definitive.

course, extremely good. If maintained, they would make the NTB an extremely successful screening test.

Results of the second study.

Given the encouraging results from the first study reported above, a second-generation test battery (version 2.0) was created. Essentially, the tests determined in version 1.1 above were all re-programmed to operate on an IBM XT or higher (or true clone). This involved creating computer versions of the Trails and Symbol-digit tests, re-programming the tracking task to operate with a joystick, and incorporating the scoring criteria into the computer so that evaluation was done automatically.

It is recognized, of course, that the new version of the battery will require different norms, and may even have a different sensitivity to the pathological subjects than the older version. Therefore, as described above, 20 of the subjects who had participated in the above study were re-tested with the new version in order to get some idea of the relationship between the two different implementations.

Analyses of variance comparing scores on versions 1.1 and 2.0 were carried out to determine which scores differed significantly. These analyses revealed that 11 out of the 32 scores were indeed different between the two test administrations. Of these differences, 5 were on tests that were significantly different in format between the two test administrations. Essentially, the data indicate that the following tests are much harder in version 2.0 than in version 1.1: Trails A, Trails B, Symbol Digit percent correct, and tracking losses. Logical reasoning reaction time was faster on version 2.0 than on version 1.1. Mathematical processing also appeared easier on version 2.0 for all variables, except that subjects got fewer correct.

This number of differences between the two versions raises the possibility that the original preliminary validation of version 1.1 may be negated. Thus, the degree to which version 2.0 was able to discriminate between pathological and non-pathological groups is also of prime interest. Results of these analyses are presented in Table V. The five test groups included pilots, non-pilot normals, substance abusers, neurologically impaired subjects, and depressives. A total of 19 out of the 32 variables (59%) significantly discriminated between the pathological and non-pathological subjects using version 2.0 of the test battery. This compares to a total of 68% significant differences for version 1.1 of the battery. Specifically, it is seen that, of the 10 tests that originally discriminated between groups in version 1.1, 7 (70%) also significantly discriminated in version 2.0. Further, 12 tests that were not significant in version 1.1 were significant in version 2.0. Thus, the basic validation of version 1.1 remains defensible for version 2.0. If anything, it appears that version 2.0 might be even more sensitive to differences among the various experimental groups.

DISCUSSION

The Neuropsychological Test Battery (NTB) described above appears to have excellent potential to answer the need for a computerized test of cognitive function that could serve as an adjunct to the routine physical examination. Its major strengths lie in the theory-based approach and in the use of a step procedure. The former offers the potential for extensive testing of all domains of higher cognitive function, while the latter provides a time- and cost-efficient screening at increasingly more diagnostic levels. The results of the initial validation studies are encouraging. Clearly, the tests selected discriminate among differing groups of normal and pathological subjects. Obviously, with further evolution of the battery, additional precision and efficiency can be added to the battery as it presently stands.

The NTB is still in the early stages of development. Although many tests are implemented clinically on the basis of less evidence, the very nature of the theory-based approach demands that far more study be carried out on the NTB before it can be validated for clinical implementation. Required studies fall into three general categories: 1) further criterion validation and cross-validation, 2) exploration of additional and alternative tests and procedures for the battery, and 3) human factors issues related to actual clinical implementation.

The first type of study is in many ways the most critical. The present studies, while establishing the validity of the basic concepts, barely scratch the surface. Cross-validating the tests and scoring criteria is an obvious first step. It would be expected that this will reveal somewhat less accurate prediction than was obtained in the training samples. Re-adjustment of criteria, re-definition of the interpretative statements, and perhaps even elimination of tests that do not cross-validate may be necessary to further refine the battery. This will interact with the second series of studies, in which continuing developments in the field of cognitive science must be monitored for identification of new resources and/or tests. It can certainly not be claimed that the multiple resources theory is complete in describing the entire domain of cognitive function. Thus, the NTB must be viewed as an evolving series of specific probes.

TABLE V. COMPARISON OF THE DIAGNOSTIC SENSITIVITY OF VERSIONS 1.1 AND 2.0 OF THE NTB AMONG THE TEST GROUPS

		SIGNIFICANCE LEVELS	
TEST	VARIABLE	VERSION 1.1	VERSION 2.0
TRAILS A	R.T.	.02	.005
TRAILS B	R.T.	.006	.008
SYMBOL-DIGIT	% CORRECT	.42	.86
	R. T.	.005	.019
TRACKING	ERROR	.019	.108
	LOSSES	.001	.213
LOGICAL REASONING			
	R. T.	.436	.012
	S. D.	.215	.028
	% CORRECT	.065	.0005
ARITHMETIC	R. T.	.073	.005
	S. D.	.029	.018
	# ATTEMPT	.046	.029
	# CORRECT	.104	.015
	% CORRECT	.801	.017
STERNBERG	SET 1 R.T.	.792	.129
	SET 2 R.T.	.411	.025
	SET 4 R.T.	.357	.007
	SET 1 S.D.	.873	.145
	SET 2 S.D.	.466	.026
	SET 4 S.D.	.500	.002
	SET 1 % COR	.680	.002
	SET 2 % COR	.260	.926
	SET 4 % COR	.703	.404
	OVERALL R.T.	.528	.016
	OVERALL %	.790	.779
	SLOPE	.082	.405
	INTERCEPT	.749	.438
ZUNG DEPRESSION	.0002	.0003	
MANIFEST ANXIETY	.027	.0005	
DYNAMIC MEMORY	R. T.	.013	.747
	S. D.	.189	.879
	% CORRECT	.088	.420

In addition, procedural or software changes might be incorporated, which could increase the diagnosticity of the battery. For example, the battery could keep track of each individual's scores over time, and automatically apply curve-fitting techniques to discern atypical patterns of change, which might provide early detection of a variety of conditions involving slow cognitive deterioration.

Finally, the third series of required studies involves making the battery appropriate for general clinical use. The effect of subject intelligence, age, sex, reading ability, motivation, etc., must be explored in more detail than has been done thus far. Instructions must be made understandable for any type of individual, and the entire battery must be human engineered so that it becomes a pleasant and self-motivating experience for everyone.

In spite of the above needs and the difficulty of the task ahead, it is important not to lose sight of the fact that a new type of routine clinical testing is embodied by this battery. Automated behavioral assessment at this level of theoretical sophistication has not been generally introduced into the routine physical examination. As noted by the AMA, current neurological screening appears increasingly inadequate in assessing the higher-level cognitive functions of interest in today's occupational environment (AMA, 1984).

The present test development suggests that it may well be possible to transfer previously expensive and complex diagnostic approaches to the screening battery, without sacrificing time or precision, by taking advantage of computerized testing and decision processes. In this sense, the NTB may be the precursor of many new test approaches which will be routinely used by examiners.

CONCLUSIONS

Based on the results of the two studies reported here, the following conclusions appear justified:

1. Age, sex, and intelligence level appear to exert moderator effects on the tests proposed for the battery, and therefore must be taken into account in any future implementations.

2. Computerized, performance-based tests are capable of achieving remarkable degrees of screening and diagnostic accuracy between normals and certain groups of subjects with diagnosed pathology.

3. A step approach to screening provides a time- and cost- efficient method of screening individuals for neurological, psychiatric, or substance abuse problems.

4. Maximum interpretative efficiency in the screening procedure can be achieved through the use of a theory-based battery of tests which probes the resources relevant to a particular real-world job or task.

RECOMMENDATIONS

In its present form, the NTB is recommended for use by experienced professionals in an experimental mode, with appropriate confirmatory testing in all cases. Under these conditions, the NTB can provide objective backup data for clinical decisions.

Recognizing that the present studies establish only the basic proof-of-concept for this approach, a series of increasingly specific and definitive studies should be carried out to permit the NTB to evolve into a stand-alone battery capable of being used in the examiner's office. Several major types of study are recommended:

1. Cross-validation studies should be carried out to establish the actual predictive accuracy of present and revised scoring algorithms. Interpretative statements should be refined in view of these studies and clinical experience.

2. Changes and additions to the basic battery should be made as the field of cognitive science progresses. Specifically, candidate tests which probe additional human resources should be studied in order to expand the applicability of the NTB to additional occupational categories.

3. In addition to the inclusion of new tests, opportunities to improve the NTB through advanced mathematical analysis of results should be explored. These include techniques for monitoring a client's performance over time, and enhanced discriminative analyses.

4. The battery should be further human engineered in such a way that it can be self-administered and automatically scored. Ultimately, feedback and appropriate follow-up recommendations should be provided directly to the client.

REFERENCES

Allport, D. A., 1980, Attention and performance. In G. Claxton (Ed.), *Cognitive Psychology: New Directions.* London: Routledge and Kegan Paul.

American Medical Association, 1984, *Review of Part 67 of the Federal Aviation Regulations and the Medical Certification of Airmen,* FAA Headquarters: Washington, D. C.

Anthony, J. C., LeResche, L., Niaz, U., von Korff, M.R., Folstein, M. F., 1982, Limits of the "Mini-Mental State" as a screening test for dementia and delirium among hospital patients. *Psychological Medicine,* 12, 397-408.

Baddeley, A. D., and Liberman, K., 1980, Spatial working memory. In R. S. Nickerson (Ed.), *Attention and Performance VIII.* Hillsdale, N.J.: Erlbaum.

Colley, A. M. and Beech, J. R. (Eds.), 1989, *Acquisition and Performance of Cognitive Skills.* John Wiley and Sons: New York.

Dick, J. P. R., Guiloff, R. J., Stewart, A., Blackstock, J., Bielawska, C., Paul, E. A., Marsden, C. D., 1984, Mini-mental State Examination in neurological patients. *Journal of Neurology and Neurosurgical Psychiatry,* 47, 496-499.

Fitts, P. M., Weinstein, M., Rappaport, M., Anderson, N., and Leonard, J. A., 1956, Stimulus correlates of visual pattern recognition: A probability approach. *Journal of Experimental Psychology,* 51, 1-11.

Folstein, M. H., Folstein, S. E., and McHugh, P. R., 1975, Mini-mental state - A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12, 189-198.

Gardner, H., 1987, *The Mind's New Science. A History of the Cognitive Revolution*, New York: Basic Books.

Goldstein, G., Tarter, R., Shelly, C., and Hegedus, A., 1983, The Pittsburgh Initial Neuropsychological Testing System (PINTS): A Neurological screening battery for psychiatric patients. *Journal of Behavioral Assessment*, 5, 227-238.

Hunter, D. R., 1975, *Development of an enlisted psychomotor / perceptual test battery.* AFHRL-TR-75-60. Air Force Human Resources Laboratory, Brooks Air Force Base, Texas.

LeRoux, C. G. J., 1988, The development of a performance battery for mental and neurological screening in the aviation medical examiner's office. Unpublished Master's Thesis, Wright State University.

Lezak, M., 1983, *Neuropsychological Assessment.* New York: Oxford University Press.

Michon, J. A., 1966, Tapping regularity as a measure of perceptual motor load. *Ergonomics*, 9, 401-412.

Moise, S. L., O'Donnell, R. D., and Hordinsky, J. P. The NTI Neuropsychological Test Battery (NTI-NTB: Hardware and Software Description, [in preparation]).

Moore, M. E. and Ross, B. M., 1963, Context effects in running memory. *Psychological Reports*, 12, 451-465.

Nelson, A., Fogel, B. S., and Faust, D., 1986, Bedside cognitive screening instruments - A critical assessment. *Journal of Nervous and Mental Disease*, 174, 73-83.

Perez, W. A., Masline, P. J., Ramsey, E. G., and Urban, K. E., 1987. *Unified Tri-Services Cognitive Performance Assessment Battery: Review and Methodology.* AAMRL-TR-87-007.

Posner, M. I., 1978, *Chronometric Exploration of Mind.* Hillsdale, N. J.: Erlbaum.

Russel, E. W., 1984, Theory and development of pattern analysis methods related to the Halstead-Reitan battery. In P. E. Logue and J. M. Shear (Eds.), *Clinical Neuropsychology: A Multi-Disciplinary Approach.* Springfield, Illinois: Charles C. Thomas.

Shingledecker, C. A., 1984, *A task battery for applied human performance research.* AFAMRL-TR-84-071. Air Force Aerospace Medical Research Laboratory, Wright Patterson Air Force Base, Ohio.

Siassi, I., 1984, Psychiatric interviews and mental status examinations. In Goldstein, G., and Hersen, M. (Eds.) *Handbook of Psychological Assessment.* New York: Pergamon.

Smith, E. E., 1968, Choice reaction time: An analysis of the major theoretical positions. *Psychological Bulletin*, 69, 77-110.

Spreen, O., and Benton, A. L., 1965, Comprehensive studies of some psychological tests for cerebral damage. *Journal of Nervous and Mental Disease*, 140, 323-333.

Sternberg, S., 1969, The discovery of processing stages: Extension of Donder's method, in: Koster, W. G. (Ed.) *Attention and Performance 2.* North-Holland, Amsterdam.

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

Tarter, R. E., and Edwards, K. L., 1986, Neuropsychological batteries. In Incagnoli, T., Goldstein, G., and Golden, C. J., (Eds.) *Clinical Application of Neuropsychological Test Batteries.* New York: Plenum.

Webster, J. S., Scott, R. R., Nunn, B., McNeer, M. F., and Varnell, N., 1984, A brief neuropsychological screening procedure that assesses left and right hemispheric function. *Journal of Clinical Psychology*, 40, 237-240.

Wickens, C. D., 1984, *Engineering Psychology.* Columbus, Ohio: Merrill.

Zung, W. W. A., 1965, A self-rating depression scale. *Archives of General Psychology*, 12, 63-70.

*U.S.GPO:1992-661-063/40037