Navy Personnel Research, Studies, and Technology Division
Bureau of Naval Personnel (NPRST/PERS-1)

Millington, TN 38055-1000 NPRST-TR-06-3 September 2006

Incremental Validity of New Computerized Aptitude Tests for Predicting Training Performance in Nine Navy Technical Schools

John H. Wolfe
San Diego, CA

Gerald E. Larson
Navy Health Research Center

David L. Alderton, Ph.D.
Navy Personnel Research, Studies, and Technology

Approved for public release; distribution is unlimited.

NPRST
research at work

NPRST-TR-06-3
September 2006

Incremental Validity of New Computerized Aptitude Tests for Predicting Training Performance in Nine Navy Technical Schools

John H. Wolfe
San Diego, CA

Gerald E. Larson
Navy Health Research Center

David L. Alderton, Ph.D.
Navy Personnel Research, Studies, and Technology

Reviewed, Approved, and Released by
David L. Alderton, Ph.D.
Director

Navy Personnel Research, Studies, and Technology
Bureau of Naval Personnel
5720 Integrity Drive
Millington, TN 38055-1000
www.nprst.navy.mil

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)* 30-11-2005	**2. REPORT TYPE** Final - Technical Report	**3. DATES COVERED** *(From - To)* Jun 1989 to Feb 1990

4. TITLE AND SUBTITLE

Incremental Validity of New Computerized Aptitude Tests for Predicting Training Performance in Nine Navy Technical Schools

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

John H. Wolfe, David L. Alderton, Gerald E. Larson

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Navy Personnel Research, Studies, and Technology (NPRST/PERS-1)
Bureau of Naval Personnel
5720 Integrity Dr.
Millington, TN 38055-1000

8. PERFORMING ORGANIZATION REPORT NUMBER

NPRST-TR-06-3

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Navy Personnel Research, Studies, and Technology (NPRST/PERS-1)
Bureau of Naval Personnel
5720 Integrity Dr.
Millington, TN 38055-1000

10. SPONSOR/MONITOR'S ACRONYM(S)

NPRST

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

A - Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

During their second week of basic training, 4,989 Navy Recruits assigned to one of nine technical training schools were administered a battery of six experimental computerized aptitude tests measuring four constructs: working memory, spatial ability, reasoning, and perceptual speed. In the afternoon of the same day, the recruits were administered the Armed Services Vocational Aptitude Battery (ASVAB). The multiple correlations of all 10 ASVAB tests with the criteria were compared with the multiple correlations of all 10 ASVAB tests plus the additional tests. Results showed that (1) the battery of new tests significantly improved the prediction of the criteria for five of the nine schools; (2) the largest validity increases (up to 16.7%) were observed for the laboratory practical performance criteria, while conventional Final School Grades (FSGs) had smaller validity increases; and (3) every new predictor significantly improved validity in one or more schools.

15. SUBJECT TERMS

Armed Services Aptitude Battery, ASVAB, Final School Grades, FSG

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Genni Arledge
UNCLASS	UNCLASS	UNCLASS	UNCLASS	17	**19b. TELEPHONE NUMBER** *(Include area code)* 901-874-2115 (882)

Foreword

This report details a study conducted between June 1989 and February 1990, where 4,989 Navy recruits were tested on an experimental computerized test battery. While research was conducted 16 years ago, these data have never been published, are quite unique and valuable, and are an important piece of the developmental history of the computer adaptive testing version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB). In the mid-1980s, the military services, prompted by advances in cognitive psychology and reductions in computer hardware costs, developed a large number of computer based cognitive ability tests. In the same timeframe, the Department of Defense, driven by advances in Item Response Theory (IRT) and reduced computer costs, embarked on a long-term effort to transform the paper-based ASVAB into an adaptive IRT test battery. These two research thrusts began coordinating toward a common implementation platform that would house CAT-ASVAB and the most promising new computer-based tests from the research laboratories in the Navy, Army, and Air Force. To determine which tests were best, two studies were commissioned. This report covers the first study, a Navy only study (often referred to as the "Navy validity study") — it was Navy only because these tests were already running on the CAT-ASVAB computer platform. The study was immediately followed by the second larger, joint-service study, the Enhanced Computer Administered Test (ECAT) battery (Alderton, Larson, & Wolfe, 1997). This Navy-only study was extensively briefed but not published at the time for a number of reasons: the immediate onset of the massive ECAT study, the departure of one of the principals (Alderton), and the churn created by the decision to close the executive-agent laboratory (Navy Personnel Research and Development Center, Base Realignment and Closure Commission, 1995). Nonetheless, even 16 years later, the results from this research are useful for scientists and it is of historical importance to document an additional facet in the march toward the implementation of CAT-ASVAB.

DAVID L. ALDERTON, Ph.D.
Director

Executive Summary

During their second week of basic training, 4,989 Navy recruits assigned to one of nine technical training schools were administered a battery of six experimental computerized aptitude tests measuring four constructs: working memory, spatial ability, reasoning, and perceptual speed. In the afternoon of the same day, the recruits were administered the Armed Services Vocational Aptitude Battery (ASVAB). The multiple correlations of all ten ASVAB tests with the criteria were compared with the multiple correlations of all ten ASVAB tests plus the additional tests. Results showed that (1) the battery of new tests significantly improved the prediction of the criteria for five of the nine schools; (2) the largest validity increases (up to 16.7%) were observed for the laboratory practical performance criteria, while conventional Final School Grades (FSG) had smaller validity increases; (3) every new predictor significantly improved validity in one or more schools.

Contents

List of Tables

List of Figures

Introduction

During the 1980s the Armed Forces' personnel laboratories conducted extensive research on new aptitude tests. This research was driven by several events. First, renewed interest (spurred in part by Congress) in predicting military job performance increased the desire for tests that were more "performance-oriented" and less scholastic in content than the Armed Services Vocational Aptitude Battery (ASVAB). Thus, attempts were made to develop new, knowledge-reduced tests of basic mental skills needed for various military occupations. Second, researchers were reacting to changes in the field of mental testing itself, including conceptual changes from psychology's "cognitive revolution" in the 1970s. Whereas traditional aptitude frameworks relied on factor analysis, cognitive frameworks emphasized information processing via mental structures and operations. With the latter approach being widely applauded as a conceptual breakthrough, test developers were anxious to apply cognitive theory and methodology to the measurement of individual differences. Third, plans to computerize the ASVAB were viewed as creating an opportunity for entirely new types of tests that exploited the computers' display, timing, and measurement capabilities. For these and other reasons, test development was a major research thrust in the 1980s. In the current paper, six experimental tests from that era are evaluated to determine their incremental validity over the ASVAB.

The ASVAB is a collection of eight power tests and two speeded tests used by all of the U.S. Armed Services for selection of enlisted personnel and qualification for admission to training schools for various military occupational specialties. Table 1 briefly describes the ASVAB tests.

Table 1
ASVAB tests and constructs (Forms 8/9/10 to FY02)

Construct	Test Name	Description
Verbal Ability	Paragraph Comprehension (PC)	A 15-item reading comprehension test: 13 minutes
	Word Knowledge (WK)	A 35-item vocabulary test using words embedded in sentences or synonyms: 11 minutes
	General Science (GS)	A 25-item knowledge test of physical and biological sciences: 11 minutes
Mathematical Ability	Arithmetic Reasoning (AR)	A 30-item arithmetic word problem test: 36 minutes
	Math Knowledge (MK)	A 25-item test of algebra, geometry, fractions, decimals, and exponents: 24 minutes
Technical Knowledge	Mechanical Comprehension (MC)	A 25-item test of mechanical and physical principles: 19 minutes
	Auto and Shop Information (AS)	A 25-item knowledge test of automobiles, shop practices, tools, and tool use: 11 minutes
	Electronic Information (EI)	A 20-item test about electronics, radio and electrical principles and information: 9 minutes
Clerical Speed	Numerical Operations (NO)	A 50-item speeded addition, subtraction, multiplication, and division test using one and two digit numbers: 3 minutes
	Coding Speed (CS)	An 84-item speeded test requiring the recognition of number strings arbitrarily associated with words in a table: 7 minutes

The ASVAB is heavily weighted towards crystallized academic skills. It lacks any direct measure of spatial ability, perceptual speed, abstract reasoning, or working memory. These and other non-crystallized abilities were assessed in the present study. The purpose of this study was to determine whether the validity of the ASVAB could be improved by supplementing it with a battery of new computerized tests of cognitive abilities.

Method

ASVAB Testing

Because ASVAB scores of record for our subjects could be more than a year old, comparison of ASVAB with new predictors could be biased if the latter were administered closer in time to the criterion. For this reason the ASVAB was readministered on the same day as the experimental tests. One-half of the subjects destined for each technical school took a computerized adaptive version of the ASVAB, while one-half took the paper-and-pencil version.[1]

The computerized adaptive test version of the ASVAB (CAT-ASVAB) has been in limited operational use since September 1990. Scores derived from the CAT-ASVAB have been equated to Form8A of the ASVAB using smoothed equipercentile equating with 8,040 applicants tested at Military Entrance Processing Stations (Segall, 1991). An empirical study of the reliability of CAT-ASVAB showed reliabilities equal to or better than the paper-and-pencil ASVAB (Moreno & Segall, 1992).

New Aptitude Tests

Each subject completed a battery of six computerized tests, presented on Hewlett-Packard Integral microcomputers operating under UNIX. All tests were written in standard C. The keyboard was modified by using a plastic mask that revealed only the designated response keys along with a key labeled HELP that could be pressed during testing to suspend the program and request assistance. The S, F, H, K, and semi-colon keys were relabeled as: A, B, C, D, and E. The space bar was relabeled ENTER. The numeric keypad keys retained their meanings. Table 2 lists the tests that were administered along with a brief description of the tests and the constructs they measure.

[1] The split design was for purposes of evaluating the cross-correlations between the paper and pencil and CAT-ASVAB tests. Few differences were found, thus the mode of administration is ignored in the remainder of this report.

Table 2
Tests and constructs in the Navy new test battery

Construct	Test Name	Description
Working Memory	Mental Counters	A 40-item working memory test using figural content
	Sequential Memory	A 35-item working memory test using numerical content
Non-verbal Reasoning	Figural Reasoning	A 35-item figural series extrapolation test
Spatial Visualization	Integrating Details	A 40-item puzzle test
	ASVAB-6 Space Perception	A 20-item paper folding test from ASVAB Form 6
Perceptual Speed	String Comparison	A 172-item reaction time based, mixed content comparison test

The following descriptions are presented in the order in which they were administered in the test battery.

Integrating Details is a complex 40-item spatial problem-solving test (Alderton, 1989). Each item consists of two separate screens. The first screen contains from 2 to 6 regular geometric puzzle pieces that must be mentally brought together to form a completed object. This is much like a jigsaw puzzle. Having connected all of the puzzle pieces, the individual must remember the final object, and then press a response key indicating that she/he is ready. Once the key is pressed, the puzzle pieces are replaced by a new screen with a single completed object. The subject must indicate if the presented object is a product of the original puzzle pieces (see Figure 1). The test score is the total number of correct responses.

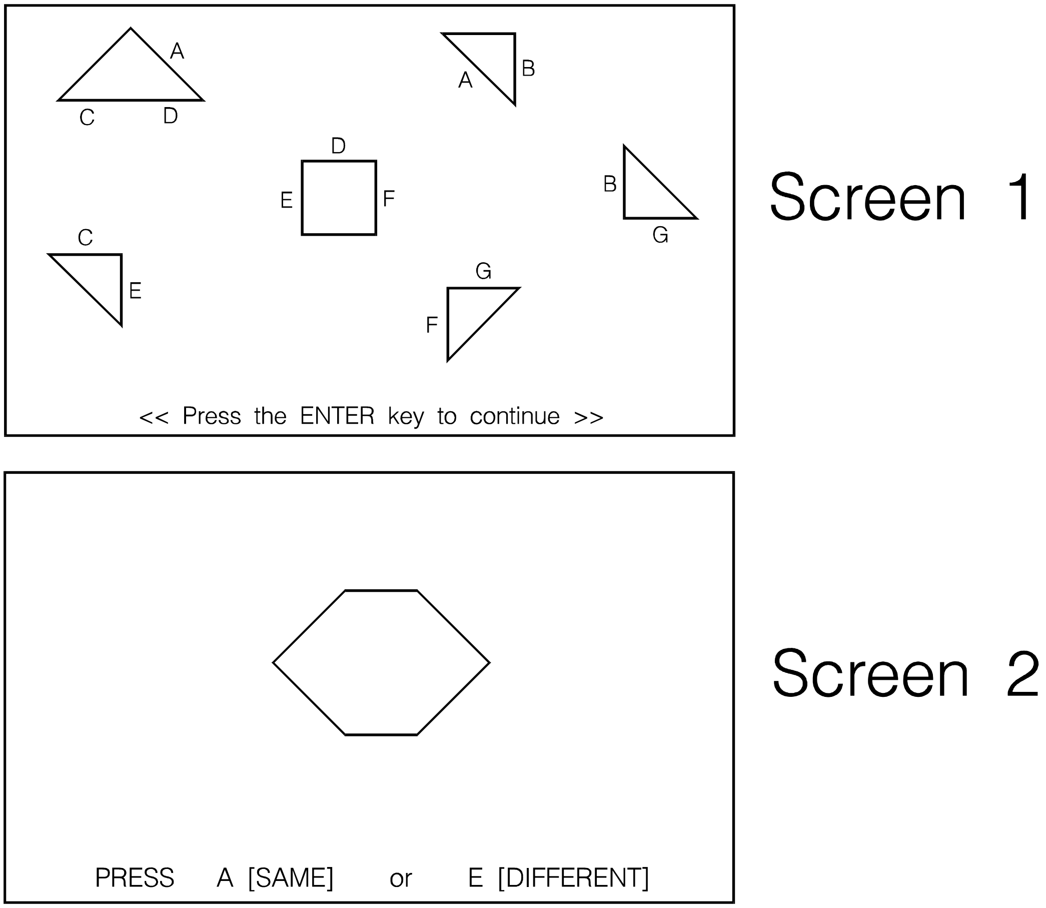
A
C
D
A
B
D
E
F
B
G
C
E
G
F
<< Press the ENTER key to continue >>
Screen 1
PRESS A [SAME] or E [DIFFERENT]
Screen 2

Figure 1. ***Integrating Details*** **facsimile item: The top frame is presented and the examinee has as long as necessary to mentally construct a complete object. Following a key press, the bottom frame is presented. The subject has as long as necessary to decide if the puzzle pieces would have constructed this object. Toggling between screens is not allowed. (Answer: Same.)**

Mental Counters (Larson, Merritt, & Williams, 1988) is a complex 40-item test of working memory, which Baddeley (1992) defines as "a brain system that provides temporary storage, and manipulation of the information necessary for such complex cognitive tasks as language comprehension, learning, and reasoning. Kyllonen and Christal (1990), Carpenter, Just, and Shell (1990), and others relate working memory to fluid intelligence. In the Mental Counters test, each screen contains three horizontal lines with each line representing an independent counter with an initial value of zero. During an item, boxes appear one at a time, either above or below one of the three lines. If a box appears above a line, the value for that counter is incremented by 1. If a box appears below a line, that counter is decremented by 1. On each trial, either 5 or 7 boxes appear at one of two rates: one every 1.33 seconds or one every .75 seconds. The subject's task is to make a series of rapid calculations and to select from four alternatives, the correct final counter values (see Figure 2). Number correct is the summary score.

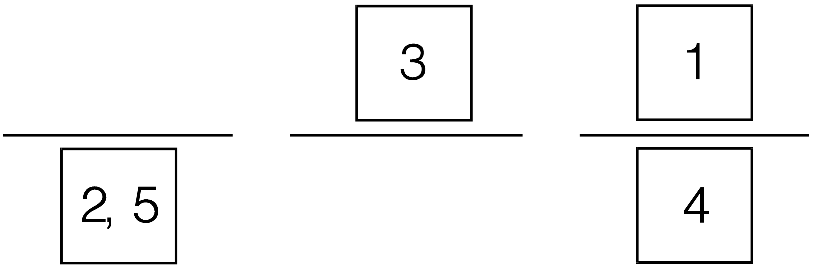

Figure 2. ***Mental Counters*** **facsimile item: Three independent counters (center horizontal lines) begin with starting values of 0. Boxes are sequentially displayed, then removed, in the order shown. If a box appears above a line the counter is incremented by 1, if below the line, it is decremented by 1. The final counter values for this item would be (in order): -2, +1, 0.**

Sequential Memory (Larson & Alderton, 1990) is another complex test of working memory. Each item begins with three to five horizontally arrayed dots on the screen. Then, each dot is assigned a numerical value that must be memorized. The item is then presented in a series of 5 or 7 "calls" to the dots; where each call is announced by briefly turning one of the dots into an "X." The person must report the digit string that corresponds to the order that the dots were called. In the second half of the test, after all the calls for an item have been made, the examinee is told to translate each number in the ordered number list into a different number and then type in the new ordered list (see Figure 3). There are 10 items in the first half of the test and 25 in the second half of the test. The score is the proportion of digits correctly reported.

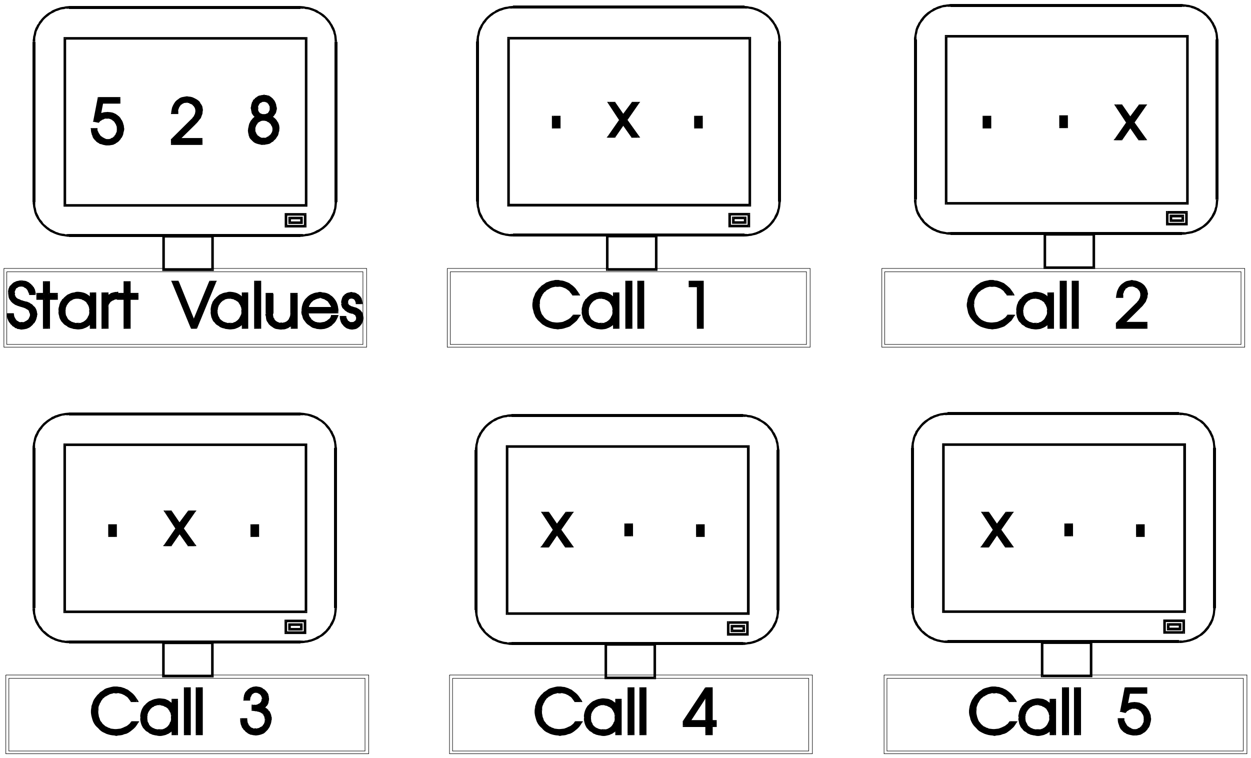

Figure 3. *Sequential Memory* facsimile item: The start values indicate the numbers assigned to each position. Following this, each time an X appears, it "calls" the corresponding number. When the X appears in the center position, the 2 is called. When the X appears in the left position, the 5 is called. When the X appears in the right position, the 8 is called. Remember the sequence of calls. (Answer: 2, 8, 2, 5, 5)

Figural Reasoning is a 30-item figural inductive reasoning (or series extrapolation) test, similar to the Cognitive Abilities Figural subtests (Hakstian & Cattell, 1976). Items use a combination of geometric forms and arbitrary figures presented in a series of four frames. The subject's task is to induce the transformation rule controlling the series and then select one of five alternatives that correctly completes the series (see Figure 4). The score is number correct in 12 minutes.

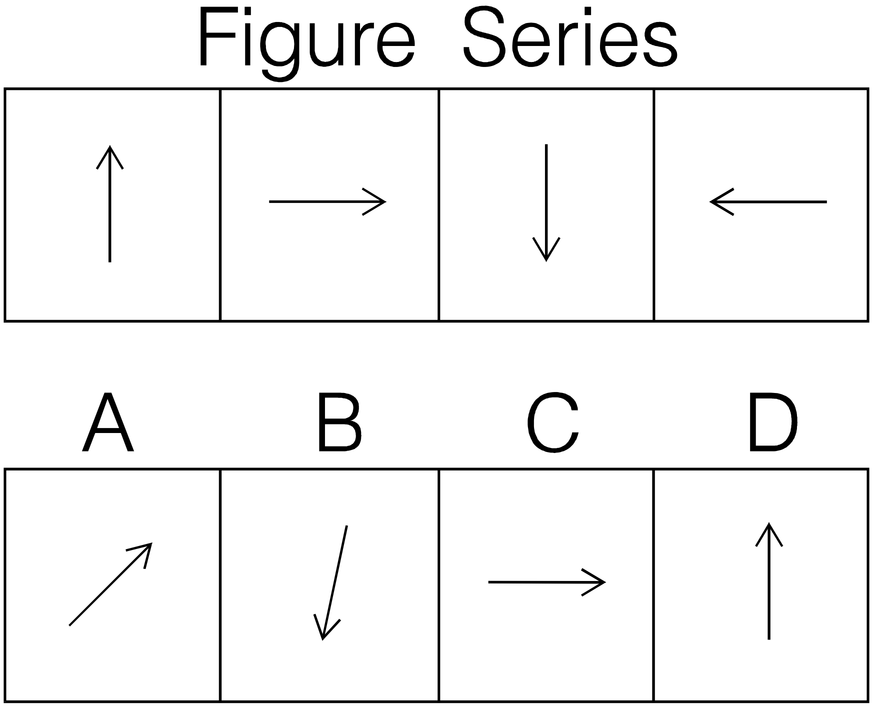

Figure 4. *Figural Reasoning* facsimile item: Which alternative shows the next frame in the figure series? (Answer: D).

Perceptual Speed (Alderton, 1990) is a clerical/perceptual speed test. Each item consists of two side-by-side symbol strings of the same length. The examinee's task is to determine whether the two symbol strings are identical, as rapidly as possible while maintaining 90 percent accuracy. Symbol string length is systematically varied from 1 to 7 elements. The test is divided into 3 subtests based on symbol type: numbers (56 items), letters (56 items), and abstract stick figures (60 items). Each item type (number of elements X symbol type) has a minimum and maximum response time bracket; if an examinee responds too quickly or too slowly she/he is warned to slow down or speed up. Cumulative accuracy is retained and used in feedback along with average response time after every 10–14 items. To control for speed/accuracy tradeoffs, the examinee is warned to slow down if accuracy drops below 85 percent or to speed up if accuracy goes above 95 percent. The primary score is the average rate score across the three subtests, where rate is defined as the proportion correct divided by the geometric mean of item response times.

658331____658331 51738690____51728690

vteqrb____vteqrb bnuoewj_____bnucewj

⍓◆□♑●♐_____⍓◆□♑●♐ ⍓□◆♐♑●_____⍓□◆♐♍●

Figure 5. *Perceptual Speed* facsimile items: Each pair of strings connected by a line must be evaluated as being the same or different. The strings consist of numbers, letters, or figures. (The above figures are a special computer font; the figures on the test were shown from a 40 character font created for the test, and each figure was a set of 2-4 connected lines.) The left column requires a "same" response and the right column requires a "different" response.

Space Perception is a computer administered version of the surface development spatial test used in ASVAB Forms 6 and 7 (discontinued in 1978). The test is similar in format and requirements to the Differential Aptitude Test battery's Space Relations subtest (Bennett, Seashore, & Wesman, 1974). Test items consist of problems showing the connected surfaces of a flat figure that can be folded into a 3-dimensional object, such as the six surfaces of an unfolded cube. The subject's task is to decide which of four alternatives correctly represents the folded version of the flat object. The score is the number correct of 20 items solved in 12 minutes.

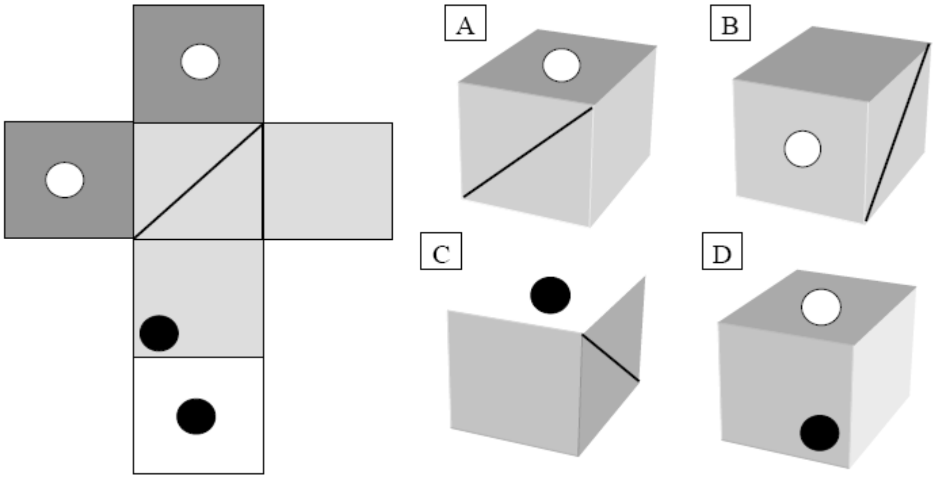

Figure 6. *Space Perception* facsimile items: Mentally fold the flat figure on the right to form a cube. Determine which of the 4 alternatives represents the correct figure. (Answer: A).

Subjects

The examinees were Navy recruits at the Great Lakes Recruit Training Center who were scheduled for technical training at one of nine Class "A" Schools. Table 3 gives the list of schools, the number of examinees expected to attend each school, the number who actually enrolled, and the number of graduates. This number was further reduced to 3,356 by eliminating cases with missing data or other factors.

Table 3
Schools for Navy validity study

Abbrev.	Title	Tested	Enrolled	Graduated
AD	Aviation Machinist's Mate	136	125	115
AMS	Aviation Structural Mechanic–Structures	122	115	104
AO	Aviation Ordnanceman	128	125	117
AV	Avionics Total, consisting of	368	330	294
	Aviation Electronics Technician	241	213	186
	Aviation Fire Control Technician	80	72	66
	Aviation Antisubmarine Warfare Technician	47	45	42
BT/MM	Boiler Tech/Machinist, consisting of	1,169	988	935
	Boiler Technician	427	353	335
	Machinist's Mate	742	635	600
GMG	Gunner's Mate–Phase I	447	427	398
HM	Hospitalman	782	832*	628
HT	Hull Maintenance Technician	454	418	391
OS	Operations Specialist	1,155	1,109	1,015
	Unassigned	228	0	0
	Total	4,989	4,469	3,997

*87 initially unassigned cases were sent to HM school.

Testing Schedule

Testing was conducted from June 1989 through February 1990. The examinees were tested in their second week of training, immediately after classification into different occupational specialties. All examinees received a maximum of three hours of experimental cognitive tests in the morning. In the afternoon, within technical school specialties, they were randomly assigned to either paper-and-pencil ASVAB (Forms 11A, 13A, or, rarely, 12A) or CAT-ASVAB (Forms 1 or 2).

Instructions

Examinees were first given written and oral information about their rights under the Privacy Act, told that the testing was for research purposes, would be kept confidential, and would have no effect on their careers, and then asked to sign a statement giving permission to be tested under these conditions.

At the beginning of each session involving computers, general keyboard familiarization instructions and practice were presented. Instructions for each of the new computer-based ability tests were presented at the beginning of each test along with several practice items. For the CAT-ASVAB, each test was preceded with instructions and several practice items. Instructions for the paper-and-pencil ASVAB tests were printed at the beginning of each subtest and read aloud by the test proctor, this included several practice problems as well.

Criteria

School performance data were obtained from a variety of sources. Final School Grades (FSG) were readily obtained from existing records. For most schools, internal consistency reliability estimates were available for FSGs. An exception was the Avionics school, where no reliability estimate could be obtained for FSG, which had a mean of 99.2 out of a possible 100, with a standard deviation of only 1.78. Avionics FSG was omitted from subsequent analyses.

In addition, every effort was made to obtain records of hands-on practical laboratory exercises. In most cases, these turned out to be simple pass-fail marks, with everyone passing. In three schools, Aviation Machinist Mate (AMS), Avionics (AV), and Hull Technician (HT), meaningful practical criteria were available. These were factor analyzed, and the factor pattern guided the construction of composites of unit-weighted criterion variables.

For the *Aviation Structural Mechanics (AMS)*, factor analysis showed two factors. A LAB criterion was defined as the sum of three practical and five performance test scores loading on the second factor. (The first factor was defined mainly by knowledge tests and because it correlated very highly with FSG, it was discarded as redundant.)

In *Avionics (AV)*, factor analysis revealed four factors, two of which were not highly correlated with FSG. The LAB1 composite was defined as the sum of six performance tests loading on one factor. The LAB2 composite was defined as the sum of eight practical exercises loading on another factor. TheLAB1 + LAB2 composite was defined as the sum of the 14 measures comprising the preceding LAB composites.

For the *Hull Technician (HT)* school, three factors were found. The QUIZ composite was the sum of 22 knowledge test scores, which correlated .67 with FSG. The LAB1 composite was the sum of 21 performance test scores from the first phase of the course. The LAB2 composite was the sum of 14 performance tests from the second phase of the course. A combined composite, LAB1 + LAB2, was defined as the sum of the 35 tests comprising LAB1 and LAB2.

Communalities were used as estimates of the reliabilities for the components of the composites, and then reliabilities of the composites were computed using the standard formulas for the correlation of sums. Table 4 shows the statistical features of the criteria. The preferred criteria, which are used later in the summary averages, are shown in boldface type and appear as the last entry for a school.

Table 4
Characteristics of school performance criteria

						Uncorrected		Corrected	
School	Criterion	N	Mean	Min	Max	Std. Dev.	r_{xx}	Std. Dev.	R_{xx}
AD	**FSG**	92	87.2	78.5	96.5	4.49	.950	6.95	.979
AMS	FSG	89	81.3	72.9	92.3	3.82	.900	6.87	.969
	LAB	89	84.3	76.7	94.8	3.75	.606	4.75	.755
AO	**LAB**	94	82.2	69.9	96.4	6.56	.880	8.32	.925
AV	LAB1	226	94.6	66.4	100.0	4.77	.512	4.90	.536
	LAB2	226	93.8	85.0	98.9	2.30	.412	2.73	.582
	LAB1+LAB2	226	94.1	81.5	99.1	2.82	.617	3.08	.678
BT/MM	**FSG**	811	86.0	75.1	99.9	5.23	.810	6.09	.860
GMG	**FSG**	324	86.0	73.0	98.7	4.79	.920	6.04	.950
HM	**FSG**	491	82.1	73.5	94.6	3.92	.930	4.84	.954
HT	FSG	322	90.5	82.6	97.4	3.11	.910	3.55	.931
	QUIZES	322	91.4	80.9	98.0	3.34	.819	4.26	.889
	LAB1	322	94.0	86.7	96.9	1.58	.788	1.68	.812
	LAB2	322	97.5	91.8	99.6	1.08	.438	1.10	.459
	LAB1+LAB2	322	95.4	90.8	97.5	1.08	.753	1.13	.775
OS	**FSG**	907	88.3	74.8	98.1	4.40	.900	5.67	.940

The corrected standard deviations of the criteria were based on the Lawley (1943) multi-variate range correction formula, using the ten pre-enlistment ASVAB tests as explicitly selected variables and the criterion as indirectly selected. The corrected reliability was computed from the formula

$$R_{xx} = 1 - \frac{s_x^2}{S_x^2}(1 - r_{xx}),$$

where r_{xx}, is the uncorrected reliability, s_x is the uncorrected standard deviation, and the corresponding corrected values are in uppercase (Gulliksen, 1950, 1987, Chapter 10, Eq. 5).

Hypothesis Testing

Since many hypotheses were tested, the Type I error associated with multiple significance tests was controlled using a hierarchical approach (Cohen & Cohen, 1983, p. 172). First, a single hypothesis for the whole study was tested, then hypotheses for each school, hypotheses for each new predictor, and finally hypotheses for school x new predictor combination.

In order to increase statistical power, the number of new predictors in most of the regression equations was reduced by combining some of them into "composites." The *Memory* composite was defined as the sum of the z-scores of Sequential Memory and Mental Counters, where the z-scores were standardized on the full sample of 4,989 Navy recruits. Similarly, the *Spatial* composite was defined as the sum of the z-scores of Integrating Details and ASVAB-6 Space Perception. These two composites plus the Perceptual Speed and Figural Reasoning tests make a set of 4 predictors in regression, which will be referred to as the 4-composite set. If a composite significantly improved prediction, its components were tested for significance later.

The multiple correlation of all ten ASVAB tests was computed for each criterion from each school. Next, the multiple correlation of the ASVAB plus four composite predictors with each criterion was computed. For each criterion, the probability associated with the difference was determined from the F-distribution with degrees of freedom equal to 4 and N - (10 + 4) - 1, where

$$F_{4,N-15} = \frac{\Delta R^2}{1 - R^2_{ASVAB+CTB}} \cdot \frac{N-15}{4}$$

These probabilities, P_i, for school i, were combined into a single number that represents the probability that the new predictors have no incremental validity in any school. For each school, only one criterion was chosen for inclusion in the aggregate probability. The combined probability is given by the chi-square distribution

$$\sum_{i=1}^{Schools} (-2\log P_i)$$

with 2 x *Schools* degrees of freedom (Fisher, 1932).

Schools: If the global null hypothesis is rejected, the previously computed probability values for each school are used to decide if the results for that school are significant.

New Predictors: If the global null hypothesis is rejected, probability values are computed for adding each new predictor alone to the ASVAB for each school. The results are accumulated across schools, using the Fisher chi-square method described above. This yields a probability value for each new predictor for the whole study. However, the probabilities for the new predictors are not independent, as they are for schools.

In a similar manner, probabilities are computed for deleting one predictor from the complete battery of ASVAB plus all new predictors. This p-value is used to test whether a given new predictor is redundant with respect to the other new predictors.

Predictor x School: If a given school and a given predictor separately show significant incremental validity, then the previously computed joint probability of using that predictor in that school is used to test the hypothesis that adding that one predictor to the ASVAB improves validity for that school.

Estimating the Magnitude of Validity Increments

All hypothesis testing was based on uncorrected correlations. To estimate the magnitude of the validity increments, several corrections were applied at various stages of the analysis. Lawley's (1943) range restriction corrections were applied to the correlation matrix of predictors and criteria, using all ten preenlistment ASVAB tests as explicitly selected variables. Multiple correlations based on either corrected or uncorrected correlations were "shrunken" to estimate population values, using the Wherry formula. If the Wherry formula yielded a negative value, it was assumed to be zero. Therefore, negative "increments" in validity were replaced by zeros. Finally, the multiple correlations were corrected for criterion unreliability, using range-corrected reliability coefficients.

For descriptive purposes, it may be useful to compute an average across all samples of the multiple Rs and their differences. Such an average, based on diverse sets of criteria and unrepresentative sampling from the domain of Navy schools, should not be given undue importance. In the present study, the multiple Rs were averaged by weighting them inversely by their asymptotic variances.[2] This procedure produces an estimate with minimum variance when the samples are drawn from the same population (Hedges, 1983). Because larger Rs produce greater weights, correlations from the same school may have different weights in the averages. The distribution of the multiple R differences may be grossly non-normal for sample sizes as large as 1,000 (Hedges, Becker, & Wolfe, 1992), and their asymptotic variances are difficult to compute in any case, so a simpler weighting method was used—they were weighted by the degrees of freedom of the larger multiple R.

All multiple correlations were Wherry-shrunken before averaging. Because negative correlation differences were replaced by zeros, the mean correlation difference may be larger than the difference in the mean correlations.

[2] Kendall and Stuart, 1979, Eq. 27.90.

Results

ASVAB Scores used for Comparison

Based on prior evidence[3], the data for the CAT and paper-and-pencil groups were combined for many of the comparisons with new predictors. Analyses of incremental validity were performed for both pre-enlistment and post-enlistment ASVAB. The incremental validities were slightly, but not significantly, lower using post-enlistment ASVAB scores, which are reported here.

Schools

Table 5 shows the uncorrected and corrected multiple correlations for each school criterion with ASVAB alone, and with ASVAB plus four new composite predictors. Where there are several criteria for a school, the one selected for subsequent analyses appears as the last entry for the school. The probabilities for the F-test of incremental validity appear in the center column. When the probabilities for the preferred criteria are combined, using the Fisher chi-square test, the probability that no validity improvement occurred in any school is less than 10^{-9}.

Significant results were obtained in five schools: Aviation Ordnance (AO), Avionics (AV), Gunner's Mate (GMG), Hull Technician (HT), and Operations Specialist (OS). OS, which had the largest sample size (N = 907) and the smallest p-value, showed a 2.1 percent increase in validity from adding four new predictors. The Boiler Technician/Machinist Mate school had the second largest sample size (N = 811) without showing any significant result. The largest percentage improvements occurred in the laboratory criteria for Avionics and Hull Technician.

Table 6 shows the averages across schools. The Fisher chi-square P value is displayed in bold in the bottom center of the table. Its value of 3.319×10^{-10} establishes that the new tests improve ASVAB validity in at least one school. A further breakdown into CAT-ASVAB and paper-and-pencil-ASVAB groups shows the same 2 percent average validity gain in each group.

[3] An earlier prototype CAT-ASVAB system using different items, algorithms, software, and computers was administered to a sample of 1,064 Navy recruits who later graduated from one of six different Navy technical training schools. School Ns ranged from 143 to 205 (Vicino & Hardwicke, 1984). No significant differences were found between the validities of CAT-ASVAB and pre-enlistment ASVAB, nor between CAT-ASVAB and an alternate form retest on paper-and-pencil ASVAB. A follow-up study with 2,054 Marine Corps, 1,487 Air Force; and 2,566 Army recruits attending 22 schools also found no significant differences in validities after corrections for range restriction (Moreno, Segall, & Kieckhaefer, 1985; Segall, Moreno, Begg, & Kieckhaefer, 1995). More recently Wolfe (1992) using the same sample described in the present paper, found no significant validity differences between CAT and paper-and-pencil versions of the ASVAB given to different subjects, nor between CAT-ASVB and pre-enlistment paper-and-pencil ASVAB for the same subjects.

Table 5
Incremental validities over post-enlistment ASVAB

			Uncorrected				Fully Corrected		
School	**Criterion**	**N**	R_{ASVAB}	$R_{ASVAB+CTB}$	**Effect***	$P(F_{4,N-15})$	R_{ASVAB}	ΔR	**% Gain**
AD	**FSG**	92	0.500	0.515	0.020	8.168 x 10^{-1}	0.796	0.000	0.0
AMS	LAB1	89	0.306	0.401	0.080	2.197 x 10^{-1}	0.639	0.017	2.6
AMS	**FSG**	89	0.404	0.417	0.013	9.116 x 10^{-1}	0.833	0.000	0.0
AO	**FSG**	94	0.554	0.643	0.182	9.727 x 10^{-3}	0.733	0.052	7.1^{***}
AV	LAB1	226	0.338	0.396	0.051	3.313 x 10^{-2}	0.423	0.061	14.5^{**}
	LAB2	226	0.383	0.450	0.070	6.326 x 10^{-3}	0.784	0.036	4.6^{***}
	LAB1+LAB2	226	0.400	0.468	0.075	4.034 x 10^{-3}	0.600	0.051	8.5^{***}
BTMM	**FSG**	811	0.494	0.498	0.006	3.498 x 10^{-1}	0.708	0.000	0.0
GMG	**FSG**	324	0.539	0.562	0.036	2.692 x 10^{-2}	0.751	0.008	1.1^{**}
HM	**FSG**	491	0.547	0.551	0.006	5.726 x 10^{-1}	0.745	0.000	0.0
HT	FSG	322	0.413	0.428	0.015	3.213 x 10^{-1}	0.602	0.002	0.3
	QUIZES	322	0.525	0.547	0.033	4.167 x 10^{-2}	0.776	0.008	1.0^{**}
	LAB1	322	0.312	0.371	0.047	6.658 x 10^{-3}	0.444	0.043	9.7^{***}
	LAB2	322	0.243	0.306	0.038	2.272 x 10^{-2}	0.355	0.059	16.7^{**}
	LAB1+LAB2	322	0.336	0.410	0.066	5.497 x 10^{-4}	0.449	0.063	13.9^{***}
OS	**FSG**	907	0.457	0.492	0.045	6.297 x 10^{-8}	0.736	0.015	2.1^{***}

*Effect = $\dfrac{\Delta R^2}{1-R_{ASVAB+CTB}}$

** p < .05

*** p < .01

Table 6
Summary of incremental validities over post-enlistment ASVAB correlations weighted inversely by their variances

		Wherry-Shrunken				Uncorrected		Fully Corrected			
Group	**N**	**Mean R_{ASVAB}**	**Mean $R_{ASVAB+CTB}$**	**Mean ΔR**	**Percent Mean**	**Mean Effect**	**Fisher** PX^2_{18}	**Mean R_{ASVAB}**	**Mean $R_{ASVAB+CTB}$**	**Mean ΔR**	**Percent Mean**
CAT	1663	0.442	0.461	0.024	5.4	0.049	1.799×10^{-4}	0.702	0.716	0.014	2.0
P&P	1693	0.474	0.505	0.023	4.9	0.048	4.652×10^{-4}	0.713	0.741	0.015	2.0
Combined	3356	0.454	0.474	0.025	5.5	0.036	3.319×10^{-10}	0.704	0.716	0.016	2.2

Note. Negative Wherry-shrunken squared multiple Rs are replaced by zeros.
Negative ΔRs are replaced by zeros.

The Fisher P values are based on the chi-square distribution of $\sum_{i=1}^{Schools}(-2 \log P_i)$ with 2 x *Schools* degrees of freedom (Fisher, 1932).

Although five of the nine schools show significant results, the largest gains occur in three schools: Aviation Ordnanceman, Avionics, and Hull Technician, where the validity increments exceed .05 correlation points. The percentage improvements in validity were 7.1, 8.5, and 13.9, respectively.

Predictors

Table 7 shows the validity increments associated with adding only one new predictor to the regression equation with all ten ASVAB tests. In keeping with our hierarchical approach, predictor results are shown only for those schools and criteria that were significant in Table 5. The nine associated probability values are not shown, but were used to generate the Fisher p-values for Table 8.

Table 8 shows mean correlations after entering one new predictor to the ASVAB or after deleting one new predictor from the combined battery of ASVAB plus new predictors. The center column of probability values to enter shows that each new predictor has significant incremental validity for at least one school. The last column shows that the Spatial composite, Mental Counters, and Sequential memory tests have unique predictive ability not measured by other tests in the battery. Either ASVAB-6 Space Perception or Integrating Details could be deleted from the battery without significant effect, but not both. Either Figural Reasoning or Perceptual Speed could be deleted from the battery without significant effect if the other tests remained.

Since all new predictors have significant incremental validity, we are justified in returning to Table 7, where we notice that the tests with significant validity increments greater than .02 occur in three schools—Aviation Ordnanceman, Avionics, and Hull Technician.[4] The two spatial tests improve validity for all three schools. Of the two working memory tests, Mental Counters increases validity in Hull Technician lab, while Sequential Memory is involved in Avionics lab. Figural Reasoning also has incremental validity in Avionics lab.

[4] The figure of .02 is arbitrarily used as a cutoff to identify the results with the greatest practical impact.

Table 7
Fully corrected incremental validities over post-enlistment ASVAB

School	Criterion	Mental Counters	Sequential Memory	ASVAB-6 Space	Integrating Details	Perceptual Speed	Figural Reasoning	Memory Composite	Spatial Composite
AO	FSG	**.000**	**.000**	**.037****	**.033***	**.003**	**.000**	**.000**	**.055****
AV	**LAB1**	**.000**	**.074****	**.000**	**.024**	**.008**	**.055****	**.045***	**.018**
	LAB2	**.001**	**.000**	**.012***	**.016***	**.018***	**.017***	**.001**	**.022****
	LAB1+LAB2	**.001**	**.026***	**.009**	**.023***	**.017***	**.038***	**.019***	**.024***
GMG	FSG	**.002**	**.001**	**.000**	**.002**	**.007****	**.001**	**.003**	**.000**
HT	**QUIZES**	**.002**	**.000**	**.000**	**.002**	**.004***	**.001**	**.002**	**.001**
	LAB1	**.039****	**.000**	**.013**	**.028****	**.002**	**.008**	**.017***	**.030****
	LAB2	**.005**	**.000**	**.038***	**.000**	**.021**	**.000**	**.000**	**.024**
	LAB1+LAB2	**.043****	**.000**	**.026****	**.028****	**.009**	**.002**	**.017***	**.042****
OS	FSG	**.009****	**.006****	**.001**	**.006****	**.000**	**.010****	**.011****	**.005****

* p < .05
** p < .01

Table 8
Fully corrected incremental validities over post-enlistment ASVAB means of correlations weighted inversely by their variances

	Simple		Entering			Deletion	P to
Predictor	Validity	$R_{ASVAB+1}$	ΔR	P to Enter	$R_{ASVAB+CTB-1}$	ΔR	Delete
Mental Counters	0.430	0.708	0.007	8.363 x 10^{-6}	0.716	0.004	0.003
Sequential Memory	0.387	0.708	0.004	4.368 x 10^{-4}	0.716	0.002	0.026
ASVAB-6 Space	0.407	0.707	0.005	1.400 x 10^{-3}	0.717	0.003	0.077
Integrating Details	0.486	0.709	0.007	3.210 x 10^{-6}	0.717	0.002	0.055
Perceptual Speed	0.202	0.706	0.003	3.830 x 10^{-2}	0.714	0.003	0.051
Figural Reasoning	0.484	0.709	0.006	1.029 x 10^{-5}	0.715	0.002	0.075
Memory Composite	0.447	0.709	0.006	1.749 x 10^{-6}	0.715	0.002	0.019
Spatial Composite	0.504	0.711	0.009	3.597 x 10^{-7}	0.712	0.006	0.001

Discussion

One fact that emerges from Table 5 is that the ASVAB is a remarkably good predictor of Navy training school grades. When corrected for restriction in range and criterion unreliability, most multiple correlations are in the mid .70s. There is little room for improvement, and the incremental validities of the new predictors are generally low for predicting grades. In contrast, laboratory performance criteria are less well predicted by the ASVAB, and here the new predictors have their greatest incremental validities. The ASVAB is best at measuring academic aptitude, or “book learning” ability, which clearly reflects the content of the test battery. Laboratory or shop work may require more fluid intelligence, spatial ability, and/or working memory, which the new predictors measure. The practical skills represented by the laboratory criteria, at least logically, may be more important for subsequent job performance than the academic learning measured with written tests. Thus the utility of new predictors for selecting personnel may be better estimated from their incremental validities for predicting labor job criteria than for school grades.

Schmidt, Hunter, and Dunn (1995) estimated that a 3 percent improvement in the average validity of the ASVAB could produce a utility increase of $83 million annually for the Navy alone. In the present study, the incremental validity conservatively averaged just over 2 percent across all schools. This increase translates into a $55 million improvement in utility for the Navy, and at least three times that for all of the military services combined. (Note that the numbers are in 1988 dollars.)

The predictors used in this study were chosen for exploratory research to determine whether the constructs of spatial ability, working memory, and perceptual speed could improve prediction of school performance. They are not necessarily the optimum enhancements to the ASVAB. The battery omits other important aspects of human performance, such as psychomotor skill. The tests themselves could be psychometrically engineered for higher reliabilities or adaptive administration. Thus further research might be able to improve the incremental validities found here, especially if laboratory or shop criteria were used.

References

Alderton, D. L., Wolfe, J. H., & Larson, J. E. (1997). The Enhanced Computer Administered Test (ECAT) Battery. *Journal of Military Psychology*, **9**, 5-37.

Alderton, D. L. (1989). *Development and evaluation of Integrating Details: A complex spatial problem solving test* (NPRDC-TR89-6). San Diego: Navy Personnel Research and Development Center.

Baddeley, A. (1992). Working memory. *Science*, 255, 556–559.

Bennett, F. K., Seashore, H. G., & Wesman, A. G. (1974). *Manual for the Differential Aptitude Tests Forms S and T* (5thed.). New York: Psychological Corporation.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices test. *Psychological Review*, 97, 404–431.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation for the behavioral sciences* (2nded.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Fisher, R. (1932). *Statistical methods for research workers* (4thed.). London: Oliver & Boyd.

Gulliksen, H. (1987). *Theory of mental tests.* Hillsdale, NJ: Lawrence Erlbaum Associates. (Originally published 1950).

Hakstian, A. R., & Cattell, R. B. (1976). *Comprehensive Ability Battery*. Champaign, IL: Institute for Personality and Ability Testing.

Hedges, L., Becker, B., & Wolfe, J. (1992). *Detecting and measuring improvements in validity* (NPRDC-TR-93-2). San Diego: Navy Personnel Research and Development Center.

Larson, C. E., & Alderton, D. L. (1990). Reaction time variability and intelligence: A “worst performance” analysis of individual differences. *Intelligence*, 14, 309–325.

Larson, C. E., Merritt, C. R., & Williams, S. E. (1988). Information processing and intelligence: Some implications of task complexity. *Intelligence*, 12, 131–147.

Kendall, M., & Stuart, A. (1979). *The advanced theory of statistics, Vol. 2* (4th ed.). New York: Macmillan Publishing Co.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity. *Intelligence*, 14, 389–433.

Lawley, D. (1943). A note on Karl Pearson’s selection formulae. *Royal Society of Edinburgh Proceedings, Section A, 62*, B 28–30.

Moreno, K., Segall, D., & Kieckhaefer, W. (1985). A validity study of the computerized adaptive testing version of the Armed Services Vocational Aptitude Battery. *Proceedings of the 27th Annual Conference of the Military Testing Association*, 29–33. San Diego, CA.

Moreno, K., & Segall, D. (1992, October). *CAT-ASVAB precision.* Paper presented at the 34th Annual Conference of the Military Testing Association, San Diego, CA.

Schmidt, F., Hunter, J., & Dunn, W. (1995). *Potential utility increases from adding new tests to the Armed Services Vocational Aptitude Battery (ASVAB) (NPRDC-TN-95-5).* San Diego: Navy Personnel Research and Development Center.

Segall, D. (1991). *Score equating development of the CAT-ASVAB.* Unpublished manuscript. Navy Personnel Research and Development Center, San Diego, CA.

Segall, D., Moreno, K., Bebb, L., & Kieckhaefer, W. (1995). *An evaluation of the validity of the experimental CAT-ASVAB.* Unpublished manuscript.

Vicino, F., & Hardwicke, S. (1984, April). *An evaluation of the utility of computerized adaptive testing.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Wolfe, J. (1992). *Validity equivalence of computerized adaptive testing and conventional administration of the Armed Services Vocational Aptitude Battery for predicting training performance in nine Navy technical schools.* Unpublished manuscript.