

AD-A252 074



DTIC

ELECTE

JUN 16 1992

J C D

CRM 91-161 / September 1991

2

An Analysis of CAT-ASVAB Scores in the Marine Corps JPM Data

D. R. Divgi
Paul W. Mayberry

DISTRIBUTION STATEMENT A

Approved for public release;
Distribution Unlimited

CNA

CENTER FOR NAVAL ANALYSES

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268

92-15394



92 6 12 054

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

Work conducted under contract N00014-91-C-0002.

This Research Memorandum represents the best opinion of CNA at the time of issue.
It does not necessarily represent the opinion of the Department of the Navy.

REPORT DOCUMENTATION PAGE

Form Approved
OPM No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources gathering and maintaining the data needed, and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Information and Regulatory Affairs, Office of Management and Budget, Washington, DC 20503.

1. AGENCY USE ONLY (Leave Blank)		2. REPORT DATE September 1991	3. REPORT TYPE AND DATES COVERED Final
4. TITLE AND SUBTITLE An Analysis of CAT-ASVAB Scores in the Marine Corps JPM Data			5. FUNDING NUMBERS C - N00014-91-C-0002 FE - 65153M PR - C0031
6. AUTHOR(S) D.R. Divgi, Paul W. Mayberry			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Center for Naval Analyses 4401 Ford Avenue Alexandria, Virginia 22302-0268			8. PERFORMING ORGANIZATION REPORT NUMBER CRM 91-161
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Commanding General Marine Corps Combat Development Command (WF 13F) Studies and Analyses Branch Quantico, Virginia 22134			10. SPONSORING/MONITORING AGENCY REPORT NUMBER
11. SUPPLEMENTARY NOTES			
12a. DISTRIBUTION/AVAILABILITY STATEMENT Approved for Public Release; Distribution Unlimited			12b. DISTRIBUTION CODE
13. ABSTRACT (Maximum 200 words) The Defense Department has developed a computerized adaptive testing (CAT) version of the Armed Services Vocational Aptitude Battery (ASVAB). During the Mechanical Maintenance phase of the Marine Corps Job Performance Measurement project, CAT-ASVAB was administered to over 1,400 Marines in Automotive Repair and Helicopter Repair occupations. The scores of these Marines were analyzed to assess the reliability of CAT-ASVAB, the potential effect of test item compromise, and how the use of computers has affected the nature of speed tests. This research memorandum presents the results of the analysis.			
14. SUBJECT TERMS Aptitude tests, ASVAB (armed services vocational aptitude battery), CAT (computerized adaptive testing), Comparison, Computer applications, JPM (job performance measurement), Performance (human), Recruits, Reliability, Scoring, Test methods			15. NUMBER OF PAGES 19
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT CPR	18. SECURITY CLASSIFICATION OF THIS PAGE CPR	19. SECURITY CLASSIFICATION OF ABSTRACT CPR	20. LIMITATION OF ABSTRACT SAR

CNA**CENTER FOR NAVAL ANALYSES**

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268 • (703) 824-2000

18 October 1991

MEMORANDUM FOR DISTRIBUTION LIST

Subj: Center for Naval Analyses Research Memorandum 91-161

Encl: (1) CNA Research Memorandum 91-161, *An Analysis of CAT-ASVAB Scores in the Marine Corps JPM Data*, by D.R. Divgi and Paul W. Mayberry, Sep 1991

1. Enclosure (1) is forwarded as a matter of possible interest.
2. The Defense Department has developed a computerized adaptive testing (CAT) version of the Armed Services Vocational Aptitude Battery (ASVAB). During the Mechanical Maintenance phase of the Marine Corps Job Performance Measurement (JPM) project, CAT-ASVAB was administered to over 1,400 Marines in Automotive Repair and Helicopter Repair occupations. The scores of these Marines were analyzed to assess the reliability of CAT-ASVAB, the potential effect of test item compromise, and how the use of computers has affected the nature of speed tests. This research memorandum presents the results of the analysis.


Lewis R. Cabe
Director

Manpower and Training Program

Distribution List:
Reverse page

Accession For	
NTIC GRAB	☒
SPAC IAB	☐
Unprocessed	☐
Justification	
By _____	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	

Subj: Center for Naval Analyses Research Memorandum 91-161

Distribution List

SNDL

A1 . DASN - MANPOWER
A1H ASSTSECNAV MRA
A2A CNR
A6 HQMC MPR & RA
Attn Code M
Attn Code MP
Attn Code MR
Attn Code MA
Attn Code MPP-54
FF38 USNA
Attn Nimitz Library
FF42 NAVPGSCOL
FF44 NAVWARCOL
Attn: E-111
FJA1 COMNAVMIIPERSCOM
FJA13 NAVPERSRANDCEN
Attn Technical Director (Code 01)
Attn Dir, Testing Systems (Code 13)
Attn Technical Library
Attn Dir, Personnel Systems (Code 12)
Attn CAT/ASVAB PMO
Attn Manpower Systems (Code 11)
FJB1 COMNAVCRUITCOM
FT1 CNET
V12 MCCDC
Attn Training and Educations Center
Attn Warfighting Center (WF-13F)

OPNAV

OP-11B

OP-136

OTHER

Military Accession Policy Working Group (18 copies)
Defense Advisory Committee on Military Personnel Testing (8 copies)
Job Performance Measurement Working Group (13 copies)

An Analysis of CAT-ASVAB Scores in the Marine Corps JPM Data

D. R. Divgi
Paul W. Mayberry

Operations and Support Division



CENTER FOR NAVAL ANALYSES

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268

ABSTRACT

The Defense Department has developed a computerized adaptive testing (CAT) version of the Armed Services Vocational Aptitude Battery (ASVAB). During the Mechanical Maintenance phase of the Marine Corps Job Performance Measurement project, CAT-ASVAB was administered to over 1,400 Marines in Automotive Repair and Helicopter Repair occupations. The scores of these Marines were analyzed to assess the reliability of CAT-ASVAB, the potential effect of test item compromise, and how the use of computers has affected the nature of speed tests. This research memorandum presents the results of the analysis.

EXECUTIVE SUMMARY

The Defense Department uses the Armed Services Vocational Aptitude Battery (ASVAB) for selection and classification of enlisted personnel. A computerized adaptive testing (CAT) version of the ASVAB has been developed by the Navy Personnel Research and Development Center. As a part of the Marine Corps Job Performance Measurement (JPM) project, CAT-ASVAB was administered to over 1,400 Marines in ground repair and helicopter repair occupations. Over 200 of the Marines took CAT-ASVAB a second time after a week or two. This research memorandum analyzes these data to address some issues related to the operational use of CAT-ASVAB.

ISSUES

A CAT-ASVAB subtest contains fewer questions than the standard paper pencil (PP) version of that subtest. Although the scores on shorter tests tend to be less precise, CAT questions are chosen to maximize the measurement precision of the test score for the person being tested, which counteracts the effect of the test being shorter. It is important to determine if CAT-ASVAB provides as much measurement precision as the PP version.

For the ASVAB to provide meaningful scores, examinees should have no prior knowledge of questions on the test. A recruiter who has come to know some ASVAB questions and tells an applicant what they are could potentially increase that applicant's score. To reduce the impact of potential compromise, an applicant is given a randomly chosen form from six operational forms of PP ASVAB. CAT has only two forms at present but, because of its adaptive nature, no two CAT tests use exactly the same set of questions. Yet, some prior knowledge of questions may help an examinee score higher. No empirical study of this issue is available.

The ASVAB contains two tests of speed: Numerical Operations (NO) and Coding Speed (CS). NO measures the rate of performing simple arithmetic operations. CS requires rapid recognition of words and numbers. Use of a computer in CAT-ASVAB has improved the accuracy of these tests because the computer can measure the time spent on individual questions. Thus, it is possible that computerized administration has changed the nature of the aptitude measured by NO and CS.

RESULTS

Based on results of the analysis in this report, it is evident that CAT-ASVAB provides more precise measurement of aptitudes than does the PP version of the test. The increase in the predictive power of the selection and classification composites is no more than 1 percent, however.

If an examinee takes CAT-ASVAB twice, scores on the second test tend to be higher if the same form of CAT is used on both the first and second tests, indicating that prior knowledge of some questions does help raise scores. Hence, it cannot be assumed that CAT provides protection against compromise of test questions.

Finally, the aptitude measured by Coding Speed is the same in CAT and PP versions of the tests. The CAT version of NO measures not only speed but mathematical aptitude as well. Although computerized adaptive testing provides more accurate measurement of aptitude, the practical value of this improvement is unclear.

CONTENTS

	Page
Introduction	1
Reliability of CAT-ASVAB Subtests	2
Susceptibility to Compromise	4
Nature of Speed Tests	5
Summary	6
References	9

INTRODUCTION

The Armed Services Vocational Aptitude Battery (ASVAB) is used for selection and classification of enlisted personnel. It contains ten subtests--General Science (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), Paragraph Comprehension (PC), Numerical Operations (NO), Coding Speed (CS), Auto and Shop Information (AS), Mathematics Knowledge (MK), Mechanical Comprehension (MC), and Electronics Information (EI). The Verbal (VE) raw score is defined as the sum of WK and PC scores. Subtests NO and CS are tests of speed in handling numerical and symbolic material. All others are power tests with liberal time limits. Standard scores rather than raw scores on the subtests are used in all decisions based on the ASVAB. Standard scores are integers from 20 to 80, with mean 50 and standard deviation 10 in the 1980 reference population [1].

Standard scores on subtests are combined into the Armed Forces Qualification Test (AFQT) score, which is the same for all services, and into occupational composites, which vary from one service to another. The AFQT is the primary score for selection. It contains subtests VE (with double weight), AR, and MK. The AFQT score may be expressed as a sum of standard scores or as a percentile rank; the former is more convenient for statistical analysis whereas the latter is the score used for selection. Composite scores are used to classify a recruit into a military occupational specialty (MOS). The Marine Corps uses four composites: (1) Mechanical Maintenance (MM), containing AR, AS, MC, and EI; (2) Clerical (CL), containing VE, MK, and CS; (3) Electronics (EL), containing GS, AR, MK, and EI; and (4) General Technical (GT), containing VE, AR, and MC. Scores on these composites have mean 100 and standard deviation 20 in the reference population.

In the computerized adaptive testing (CAT) version of the ASVAB, there is a large pool of items for each subtest. Different items are administered to different examinees in an attempt to maximize measurement precision for each examinee. The adaptive strategy of item selection consists of using the current estimate of the examinee's ability to select the next item. The next item is chosen to maximize the information it will provide about the examinee's ability. As a result of this procedure, a CAT subtest of any given length is more reliable than a PP version containing the same number of items. Using the results obtained in equating studies, CAT scores are expressed as raw number correct scores equivalent to those on PP Form 8a. These are converted to standard scores using the same table as the one used with Form 8a. Tables for converting CAT ability scores to raw scores were provided by the Navy Personnel Research and Development Center.

The Marine Corps is conducting a multiphase Job Performance Measurement (JPM) project. In the Mechanical Maintenance (MM) phase of this project, CAT-ASVAB was administered to over 1,400 Marines in Automotive Repair and Helicopter Repair occupations. If a Marine's CAT scores were higher than his or her previous ASVAB scores, the CAT scores

became the scores of record. The chance to improve the scores of record motivated the Marines to perform well on CAT. The CAT-ASVAB was re-administered to about 200 randomly selected Marines a week or two later. This subsample yields reliabilities of CAT-ASVAB subtests and composites. The total sample is useful for other analyses. Data on each Marine included ASVAB standard scores on the test taken for enlistment. Because the number of women in the study was small, and because there is evidence that the CAT version tends to underestimate their aptitude in Auto-Shop Information, women were excluded from all analyses.

RELIABILITY OF CAT-ASVAB SUBTESTS

Each CAT-ASVAB subtest contains fewer items than its PP version. Shorter tests provide less reliable measures of performance than longer ones, other things being equal. On the other hand, the adaptive nature of CAT makes it more reliable than a PP test of the same length, if the average item in the CAT pool is as good as the average item in the PP version. The empirical question is: given the test lengths and item pools of CAT-ASVAB, how do the reliabilities of its subtests and composites compare with those in the PP version?

The sample available for reliability analyses consisted of 202 Marines who were administered the CAT test twice. Because of policy constraints, scores from the second test could not be used to improve an individual's scores of record. Therefore, some examinees may have been less motivated on the second test, indicating that some data editing was needed to remove possibly unmotivated examinees. In each administration, scores on each power subtest were standardized to mean 50 and standard deviation 10 in the sample. Denote these by Z_{s1} and Z_{s2} , where subscript s indicates the subtest and 1, 2 represent first and second administrations of the test. For each Marine, two indices were created from these standardized scores. The first index, based on the total score on all power subtests, was given by

$$Q_1 = \left(\sum_s Z_{s1} - \sum_s Z_{s2} \right)^2 .$$

The second index was

$$Q_2 = \sum_s (Z_{s1} - Z_{s2})^2 ,$$

where each sum was taken over all eight power subtests. Distributions of both indices were examined, and cutoffs set near the 95th percentile. The cutoff was 90 for Q_1 and 210 for Q_2 . Marines were retained for analysis if both indices were below their cutoffs, which led to a useful sample of 188 persons.

There are two equivalent forms of CAT-ASVAB. The CAT form used on the retest was chosen at random, without attention to which form was used in the original test. As a result, 91 Marines got the same form on retest as on the initial one, and 97 got a different form. Only the latter subsample was used to calculate reliabilities.

An estimate of reliability in the recruit population is given by the correlation of a subtest from initial test to retest on a different form. These reliabilities have little operational significance because recruits have already been selected on the basis of their scores on the enlistment ASVAB. To decide how well CAT will work in selection and classification, reliabilities in the unrestricted national population are needed. The conversion of statistics from recruits to the national population is termed "correction for range restriction." To make such a correction, it is assumed that the error variance of any score, on a subtest or a composite, is the same in both populations. Variances of standard scores in the national population are known. Therefore, it is easy to compute corrected reliabilities. Take GS as an example. Its variances in the recruit sample and the national population are 35.0 and 100; its reliability in the sample is 0.739. Hence, its error variance is $EVAR(GS) = 35.0 (1 - 0.739) = 9.14$, and its corrected reliability is $1 - 9.14/100 = 0.909$. The same method applies to composites as well, except that the error variance of a composite is the sum of error variances of subtests in the composite. Take AFQT as an example. The AFQT sum of standard scores is $AFQT = 2 VE + AR + MK$, and hence, its error variance is

$$\begin{aligned} EVAR(AFQT) &= 4 EVAR(VE) + EVAR(AR) + EVAR(MK) \\ &= 4 (2.6) + 8.3 + 9.2 = 27.9 \end{aligned}$$

Variance of AFQT in the national population is 1,321.2, and the corrected reliability is 0.979.

Reliabilities of CAT-ASVAB scores were compared with those of PP scores. Alternate form reliabilities of PP subtests are available in table 6 of the technical supplement to the counselor's manual for the Student Testing Program [2]. Composite reliabilities were computed from these subtest reliabilities (as was done for CAT).

Table 1 contains the sample standard deviations (on the initial test), sample reliabilities, error variances, corrected reliabilities of CAT subtests and composites, and corrected reliabilities in the PP version. Sample standard deviations and reliabilities are not reported for composites because the range-corrected reliabilities were computed directly from subtest error variances. The largest fractional increase in composite reliability is 2.6 percent for MM.

Assuming that CAT and PP scores measure the same trait, predictive validity is proportional to the square root of reliability. Hence, the percent increase in validity is about half that in reliability. Therefore, according to composite reliabilities in table 1, an increase

in reliability raises validity in the population by no more than the 1.3 percent increase for MM. The small size of this gain is caused by the fact that composite reliabilities of PP ASVAB are already close to the maximum possible value of 1.

Table 1. Sample and corrected statistics for CAT scores and corrected reliabilities of PP scores

Score	CAT-ASVAB			PP ASVAB	
	Sample standard deviation	Sample reliability	Error variance	Corrected reliability	Corrected reliability
GS	5.9	.739	9.1	.909	.830
AR	6.0	.770	8.3	.917	.900
WK	3.8	.773	3.4	.966	.920
PC	5.9	.713	9.8	.902	.710
NO	6.6	.794	9.0	.910	.820
CS	6.7	.757	10.8	.892	.840
AS	4.7	.833	3.6	.964	.860
MK	7.2	.821	9.2	.908	.870
MC	6.5	.720	12.0	.880	.820
EI	5.7	.697	9.9	.901	.780
VE	4.0	.837	2.6	.974	.930
AFQT			27.9	.979	.961
MM			33.8	.972	.948
CL			22.6	.965	.945
EL			36.5	.971	.950
GT			22.9	.967	.950

Based on these results, the following summary statements can be made:

- o CAT scores are more reliable than PP scores.
- o The ratio of CAT to PP reliability is higher for subtests than for AFQT and the composites. In the latter group, the largest percent increase from PP to CAT reliability, observed for MM, is 2.5 percent.
- o Composite validities can be expected to increase by only about 1 percent if the PP version is replaced by CAT.

SUSCEPTIBILITY TO COMPROMISE

As mentioned above, 91 Marines were tested with the same form on the initial and second tests. The results presented below indicate that some items used in the initial test were repeated on the retest. If the examinees tended to answer these questions the same way on the retest as on the initial test, this commonality makes the test-retest correlation higher than the correlation that appears when different forms were used on the test and the retest. Such a difference in test-retest correlations was found; the average difference over the 11 subtests was 0.08. For this reason, Marines who had the same CAT form on both tests were excluded from reliability analyses.

If some or all of the Marines thought about the test questions and learned the correct answers to some items they had answered wrong on the first test, retest scores would increase. Mean scores were analyzed to determine if this phenomenon had occurred. To simplify the analysis, standard scores on power tests (except VE) were added up and only this sum was analyzed. The sum on the initial test was subtracted from the retest. The mean change was -4.60 points when initial and retest forms were different; that is, the mean score went down, which is consistent with the fact that, on retest, there was no incentive to score high. The mean change was 1.13 points when the forms were the same. The difference between the means is statistically significant at the 0.01 level. Thus, Marines tended to score higher on the retest when the form was the same as on the initial test. This shows the effect of prior exposure to some of the items, even in the absence of any coaching.

NATURE OF SPEED TESTS

As seen in table 1, CAT versions of speed tests are more reliable than PP versions. The next question is whether the CAT versions measure anything other than speed. All 1,434 Marines were included in the analyses of this question, and only the initial test scores were used.

The ASVAB measures four factors: speed, verbal, mathematical, and technical abilities [3]. Scores on these factors, using standard scores, were defined as follows: $VERB = GS + 2 VE$, $MATH = AR + MK$, and $TECH = AS + MC + EI$. For each Marine, available scores included PP scores from the enlistment ASVAB as well as the CAT scores.

CAT NO was regressed on CAT CS, PP NO, and PP CS. Similarly, the other speed tests were regressed on the remaining speed scores. Then, in all regressions, the factor scores VERB, MATH, and TECH, in the same battery as the dependent variable, were added to the list of predictors. Table 2 presents the sum of squares explained by each predictor when it is added to the equation. (It should be remembered that the order of entering speed tests into the regression varies from one dependent variable to another.) The squared multiple correlation

(adjusted to make it an unbiased estimate) is reported for regression on speed tests only, and after the other factors have been added to the equation.

Table 2. Explained sums of squares and squared multiple correlations in predicting scores on speed tests

	Dependent variable			
	CAT		PP	
	NO	CS	NO	CS
SS(CAT NO)		13,369	6,453	4
SS(CAT CS)	14,399		0	5,399
SS(PP NO)	8,277	0		14,293
SS(PP CS)	4	5,835	13,352	
RSQ	0.338	0.308	0.354	0.329
SS(VERB)	970	581	381	110
SS(MATH)	4,120	313	66	11
SS(TECH)	873	14	146	196
RSQ	0.427	0.322	0.363	0.333

First, consider the simpler case of CS. After regressing on three speed subtests, the squared multiple correlation, adjusted for capitalizing on chance, was 0.308. After adding the three factor scores, the squared multiple correlation increased to only 0.322. Thus, the CAT version of Coding Speed appears to measure little other than speed of symbol recognition. For NO, the adjusted R-square using only the speed subtests was 0.338. After adding the other factors, this increased to 0.427. The MATH factor was responsible for 69 percent of the additional variance explained. Thus, the CAT version of Numerical Operations measures speed and, to a smaller extent, mathematical aptitude. Like CAT CS, the PP versions are also almost pure measures of speed.

SUMMARY

This report addresses three issues concerning the computerized adaptive version of the ASVAB: reliability, the potential effect of item compromise, and the nature of speed subtests. Based on the analysis, CAT-ASVAB is more reliable than the PP version. For composite scores that are the level on which selection and classification decisions are made, however, the increase is only about 2 percentage points. This reflects an increase in composite validity of about 1 percent. In translating increased reliability into increased validity, it is assumed that the two versions measure exactly the same

trait. The analysis indicates that this assumption does not hold for Numerical Operations. Any departures from the assumption can yield actual validity above or below the calculated value. At present, there is too much uncertainty in estimates of reliabilities and validities to make any strong statement. Based on the available evidence, it appears that an increase in predictive validity, resulting from higher reliability, is slight.

The CAT-ASVAB cost benefit analysis [4] claims that the CAT version provides greater protection against test compromise because it "makes item specific training of examinees, or the knowledge of individual test items in advance, essentially useless. (p. E-4)" Such is not the case. CNA's analysis of mean scores shows that retest scores were higher on average when the form was the same as that used on the initial test. Thus, despite the absence of coaching or even of any pressing reason to score high, prior knowledge of some items tended to raise scores.

The last analysis showed that the CAT version of Numerical Operations measures not only speed in making simple calculations, but mathematical aptitude as well. This may be due to the fact that, in the CAT version, questions are answered by pressing keys, which takes less time than filling spaces on multiple choice answer sheets. Thus, more of the time spent on an item is actually used for the numerical operation involved. If the trait measured by the CAT version of a subtest differs from that measured by the PP version, standard formulas that relate predictive validity to reliability cannot be used.

REFERENCES

- [1] CNA Report 116, *The ASVAB Score Scales: 1980 and World War II*, by Milton H. Maier and William H. Sims, Jul 1986
- [2] U.S. Military Entrance Processing Command, *ASVAB: Technical Supplement to the Counselor's Manual*, Jul 1985
- [3] CNA Memorandum 83-135, *A Factor Analysis of ASVAB Form 8a in the 1980 DOD Reference Population*, by Peter H. Stoloff, Aug 1983
- [4] Automated Sciences Group and CACI, INC.-FEDERAL, *CAT-ASVAB Program: Concept of Operation and Cost/Benefit Analysis*, Mar 1988