AD-A263 994

DTIC
S ELECTE
MAY 1 1 1993
C D

Sensitivity and Fairness of the Marine Corps Mechanical Maintenance Composite

D. R. Divgi
Paul W. Mayberry
Neil B. Carey

50 Years
CNA 1992
CENTER FOR NAVAL ANALYSES
4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268

93 5 03 20 1

93-09894

REPORT DOCUMENTATION PAGE

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	3. REPORT TYPE AND DATES COVERED
	June 1992	Final

4. TITLE AND SUBTITLE	5. FUNDING NUMBERS
Sensitivity and Fairness of the Marine Corps Mechanical Maintenance Composite	C - N00014-91-C-0002 PE - 65153M
6. AUTHOR(S) D.R. Divgi, Paul W. Mayberry, Neil B. Carey	PR - C0031

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Center for Naval Analyses 4401 Ford Avenue Alexandria, Virginia 22302-0268	CRM 92-71

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSORING/MONITORING AGENCY REPORT NUMBER
Commanding General Marine Corps Combat Development Command (WF 13F) Studies and Analyses Branch Quantico, Virginia 22134	

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT	12b. DISTRIBUTION CODE
Approved for Public Release; Distribution Unlimited	

13. ABSTRACT *(Maximum 200 words)*

A score used for selection of classification should predict the performance of different population subgroups equally well. This research memorandum analyzes the prediction of hands-on performance in the Automotive Mechanic specialty, using the Marine Corps' Mechanical Maintenance (MM) composite

14. SUBJECT TERMS	15. NUMBER OF PAGES
Aptitude tests, JPM (Job performance measurement), Mathematical prediction, Performance (human), Performance tests, Regression analysis, Scoring, Validation	16
	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

22 June 1992

MEMORANDUM FOR DISTRIBUTION LIST

Subj: CNA Research Memorandum 92-71

Encl: (1)   CNA Research Memorandum 92-71, *Sensitivity and Fairness of the Marine Corps Mechanical Maintenance Composite*, by D. R. Divgi, Paul W. Mayberry, and Neil B. Carey, Jun 1992

1.   Enclosure (1) is forwarded as a matter of possible interest.

2.   A score used for selection or classification should predict the performance of different population subgroups equally well.   This research memorandum analyzes the prediction of hands-on performance in the Automotive Mechanic specialty, using the Marine Corps' Mechanical Maintenance (MM) composite.

Donald J. Cymrot
Director
Manpower and Training Program

Distribution List:
Reverse page

Subj: Center for Naval Analyses Research Memorandum 92-71

Distribution List

SNDL
A1   DASN - MANPOWER
A1H   ASSTSECNAV MRA
A2A   CNR
A5   PERS-11B
A5   PERS-23
A6   HQMC MPR & RA
            Attn   Code M
            Attn   Code MP
            Attn   Code MR
            Attn   Code MA (2 copies)
            Attn   Code MPP-54
FF38   USNA
            Attn   Nimitz Library
FF42   NAVPGSCOL
FF44   NAVWARCOL
            Attn: E-111
FJA1   COMNAVMILPERSCOM
FJA13   NAVPERSRANDCEN
            Attn   Technical Director (Code 01)
            Attn   Technical Library
            Attn   Dir, Personnel Systems (Code 12)
            Attn   Manpower Systems (Code 11)
FJB1   COMNAVCRUITCOM
FT1   CNET
V12   MCCDC
            Attn   Training and Educations Center
            Attn   Warfighting Center
            Attn   Studies and Analyses Branch

OTHER
Military Accession Policy Working Group (17 copies)
Defense Advisory Committee on Military Personnel Testing (8 copies)
Joint Service Job Performance Measurement Working Group (12 copies)

Sensitivity and Fairness of the Marine Corps Mechanical Maintenance Composite

D. R. Divgi
Paul W. Mayberry
Neil B. Carey

Operations and Support Division

ABSTRACT

A score used for selection or classi-
fication should predict the performance of
different population subgroups equally
well. This research memorandum analyzes
the prediction of hands-on performance in
the Automotive Mechanic specialty, using
the Marine Corps' Mechanical Maintenance
(MM) composite.

EXECUTIVE SUMMARY

The Armed Services Vocational Aptitude Battery (ASVAB) is used to select and classify enlisted personnel. The Armed Forces Qualification Test is used to select personnel, and the service composites are used to classify them into occupational specialties. The Marine Corps uses the Mechanical Maintenance (MM) composite for classifying personnel into occupations involving maintenance and repair of mechanical systems.

In a recent report, the General Accounting Office (GAO) raised some questions about the fairness of composite scores used by services for technical occupations. GAO has concluded that composites are less successful in predicting performance of women and minorities than they are in predicting that of white males, especially if performance is measured in the field. DOD is preparing a response to the GAO report, based on analyses of data from a large number of occupations from the four services.

In the MM phase of its Job Performance Measurement (JPM) project, the Marine Corps has developed hands-on performance tests (HOPTs) for the Automotive Mechanic specialty (MOS 3521) and four helicopter repair specialties (MOSs 6112 to 6115). The content of each test was based on extensive job analysis based on the Individual Training Standards. Each test was scored by former Marines who had experience in the occupation and who had been trained to score performance as objectively as possible. A report of the National Academy of Sciences calls a test score obtained in this manner "the benchmark measure" of job performance.

This study analyzes only the Automotive Mechanic data because sample sizes in the others were too small for useful analysis. Even this occupation had few women and Hispanics, and therefore only blacks and whites were compared. After removing cases with incomplete data, the sample contained 118 blacks and 632 whites.

Fairness of the MM composite means that a specific MM score predicts the same HOPT score for all individuals, regardless of their group membership. This similarity of predicted scores is tested via regression analysis in which the slopes of the prediction equations for the two groups are compared, and so are the intercepts. The hypothesis of fairness was tested separately for two MM scores: one used for enlistment in the Marine Corps, and the other from an ASVAB administered concurrently with the HOPT as part of the JPM project. For both sources of aptitude information, differences between blacks and whites in the slopes and intercepts of the regression lines were found to be statistically nonsignificant.

In summary, the evidence indicates that the MM composite score is equally sensitive for both subgroups as a predictor of hands-on performance on the job. In addition, it does not underpredict or overpredict the performance of either subgroup.

CONTENTS

INTRODUCTION

The Armed Services Vocational Aptitude Battery (ASVAB) is used to select and classify 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 a mean of 50 and a standard deviation of 10 in the 1980 reference population.

Standard scores from certain subtests are combined to compute an individual's Armed Forces Qualification Test (AFQT) score, which is the primary score used to select individuals for military service. Composite scores are used within each service to classify a recruit into a military occupational specialty (MOS). The Marine Corps uses four composites: Mechanical Maintenance (MM), which contains AR, AS, MC and EI; Clerical (CL), which contains VE, MK, and CS; Electronics (EL), which contains GS, AR, MK, and EI; and General Technical (GT), which contains VE, AR, and MC. Scores on these composites have a mean of 100 and a standard deviation of 20 in the reference population.

The General Accounting Office (GAO) has raised some questions about the fairness of service composites used for technical specialties [1]. According to the Executive Summary of GAO's report,

> GAO concluded that, for most recruits, the services' selection criteria are moderately successful at predicting individual performance during classroom technical training. However, they are notably less successful for women and minority recruits.... Only the Army systematically collects data on the field performance of individual graduates in a way that would allow comparison of a graduate's on-the-job performance with his or her entry level ability and classroom performance. These data reveal an even weaker connection for women and minority group members between criteria used to assign them to technical specialties and their later field performance [1, p. 3].

Fairness means that a score used for selection or classification predicts the same performance level for all individuals with the same score, regardless of their group membership. This similarity of predicted scores is tested via regression analysis, in which the slopes of the prediction equations for the two groups are compared and then, if

-1-

the difference is nonsignificant, intercepts are compared. Comparability of slopes from the separate regressions for each group implies equal sensitivity of predictors Equality of intercepts indicates that the test does not underpredict or overpredict the performance of any group. These hypotheses were evaluated by using the MM score to predict scores on a hands-on performance test (HOPT) that measured proficiency on representative job tasks. The MM composite is used for occupations involving mechanical repair and maintenance.

Figure 1 shows the regression lines for a test that is fair to groups A and B. The comparison of the slopes determines whether the regression lines are parallel. The second significance test determines whether the regression intercepts are significantly different. A fair test is one in which the slopes and intercepts for the two groups do not differ and hence the lines overlap. Therefore, all aptitude scores result in equal predicted scores for the two groups $[\hat{Y}(A) - \hat{Y}(B)]$.
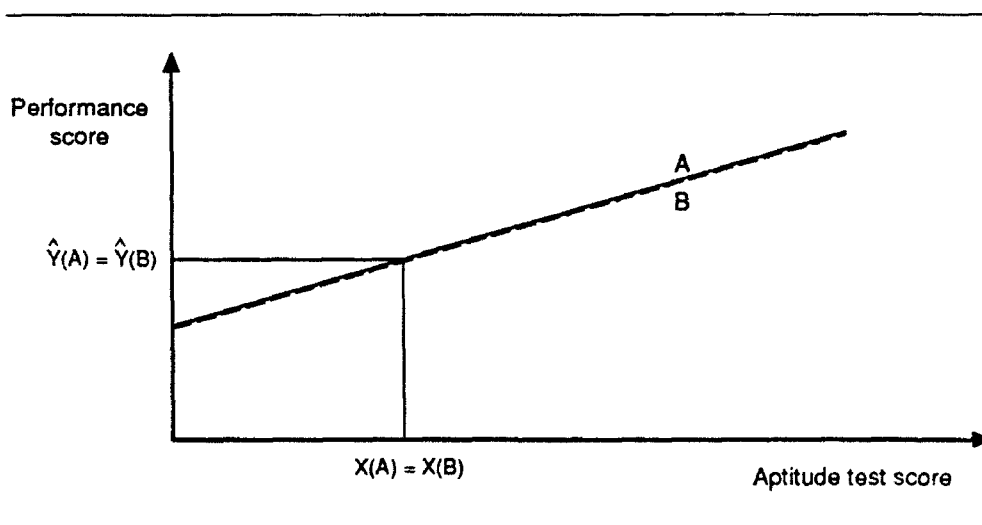

Figure 1. Regression lines for a fair test: equal slopes and intercepts

DATA

In the Mechanical Maintenance phase of its Job Performance Measurement (JPM) project, the Marine Corps developed HOPTs for five occupations for which MM is used as the classification composite. These are the Automotive Mechanic specialty (MOS 3521) and four helicopter specialties (CH-46, MOS 6112; CH-53A/D, MOS 6113; UH/AH, MOS 6114; and CH-53E, MOS 6115). Each test consists of a sample of tasks that a mechanic in that specialty needs to perform in the course of his or her work. Requirements of each job were determined using the Individual Training Standards of the Marine Corps. Each task was divided into a number of steps, each

of which was scored as performed correctly or not. The test was adminis-
tered by former Marines with relevant job experience. The administrators
were trained to score performance as objectively as possible [2]. A
score resulting from such a process has been referred to as the "bench-
mark measure" of job performance [3, p. 95].

A data set was constructed for each MOS containing the cases for
which a valid HOPT score was available. The largest total sample size
among helicopter MOSs was 215 for MOS 6114; the largest minority sample
size was 22; and no women were in the MOSs. Therefore, analyses were
performed only for the Automotive Mechanic specialty.

Time in service (TIS) exceeded 10 years in only four cases, with
values ranging from 136 to 160 months; these cases were excluded as
outliers (i.e., cases that are unusually far from most Marines). The
available ASVAB scores are those with which the Marine enlisted, and
scores from a computerized adaptive testing (CAT) version of the ASVAB
that was administered concurrently with the HOPT. All cases with missing
enlistment or CAT scores were deleted. The remaining sample contained
only 44 women and 83 Hispanics (the latter number being distinctly
smaller than the sample size for blacks). These groups were excluded
from the study because the sample contained too few of them for useful
analysis. The final sample, with complete data for each Marine, con-
tained 118 blacks and 632 whites.

ANALYSES AND RESULTS

TIS is a powerful predictor of hands-on performance. That is,
given equal ASVAB scores, senior Marines score higher on the average
than junior ones due to on-the-job training. The rate of growth slows
as time increases. Therefore, TIS and its square were included as
predictors along with MM scores.

In simple regression analyses, outliers are usually removed. In
multiple regressions, however, this simple approach can be inadequate.
Each case may need to be examined in terms of how it affects the esti-
mates of the the regression weights. The effect is quantified as fol-
lows: The weights are estimated using the entire sample. Then they are
recomputed with one observation removed. For each predictor, the latter
estimate is subtracted from the former, and the difference is divided by
the standard error of the estimate [4]. The ratio yields the "influ-
ence" of the observation on the estimated coefficient of the predictor.
A large value of either sign shows that the observation changes the
estimate substantially, and thus behaves like an outlier in simple
regression.

As the minority sample size was only 118, a few influential cases
could affect the result substantially. Therefore, each significance
test was preceded by influence analysis. Cases with extreme values of
the influence function were excluded, and then a significance test was
performed on the edited sample.

Specifically, let us consider analysis of the MM score used for enlistment. The regression equation initially included a term to represent the difference between slopes for blacks and those for whites. Influence on this term was calculated for all individuals in the sample. The standard deviation of th influence values was .038, and the mean was zero as expected. Four examinees with influence exceeding .25 in magnitude were deleted from the sample. Using the edited sample, the F ratio for difference between slopes was 0.54, which is statistically nonsignificant. Therefore, in he analysis of difference between intercepts, slopes in the two groups were set to be equal. Then influence analysis was performed for difference between intercepts. Standard deviation of influence values was .041. Again, cases with influence above .25 in magnitude were deleted. This further reduced the sample size by three. The F ratio for difference between intercepts was 3.62, which is not significant at the .05 level.

A similar procedure was followed with the MM score from concurrent CAT-ASVAB. The cutoff value for size of influence was again .25. Three cases were deleted for the analysis of slopes and two more for the analysis of intercepts. Table 1 presents detailed results for enlistment and CAT-ASVAB.

Table 1. Analyses of regression slopes and intercepts using enlistment and CAT-ASVAB scores

	Enlistment		CAT	
	Blacks	Whites	Blacks	Whites
Slope				
Sample sizes	114	632	115	632
Estimates	.22	.31	.38	.35
F ratio	0.54		0.17	
Significance level	.46		.68	
Intercept				
Sample sizes	111	632	114	631
Estimates	37.67	39.15	32.70	34.09
F ratio	3.62		3.58	
Significance level	.057		.059	

Most cases excluded due to extreme-influence values were blacks. Because the black sample size is less than a fifth of the white sample size, a black individual tends to influence the difference between subgroups more than a white individual.

DISCUSSION

The statistical significance of the intercept differences is even weaker than it appears. Since four F tests have been performed, a .05 significance level for the entire set of tests requires that, for an individual F ratio to be considered significant, its tail probability should be smaller than $.05/4 = .0125$. If the .05 significance level is applied to individual F tests, the overall significance level is $.05*4 = .20$. Thus, the set of four F tests reported above is nonsignificant at the .20 level.

In summary, Marine Corps JPM results for the Automotive Mechanic specialty, using the hands-on performance test as the criterion, show that the Mechanical Maintenance composite is equally sensitive for blacks and whites. The results also show that the regression equation does not overpredict or underpredict the performance of blacks.

REFERENCES

[1]   United States General Accounting Office, *Military Training: Its Effectiveness for Technical Specialties is Unknown.* Washington, DC: Government Printing Office, Oct 1990

[2]   CNA Research Memorandum 91-242, *Development and Scoring of Hands-On Performance Tests for Mechanical Maintenance Specialties*, by Neil B. Carey and Paul W. Mayberry, Mar 1992

[3]   Alexandra K. Wigdor & Bert F. Green, Jr., Eds. *Assessing the Performance of Enlisted Personnel: Evaluation of a Joint Service Research Project.* Washington, DC: National Academy Press, 1986

[4]   D. A. Belsley, E. Kuh, & R. E. Welsh. *Regression Diagnostics.* New York: John Wiley & Sons, 1980