AD NUMBER: ADB118782
LIMITATION CHANGES
TO: Approved for public release; distribution is unlimited.
FROM: Distribution authorized to US Government Agencies only; Administrative/Operational Use; 9 Mar 1988. Other requests shall be referred to Deputy Chief of Staff (RD&S), Headquarters, Marine Corps Washington, DC, 20380.
AUTHORITY
CNA ltr dtd 15 Dec 1988

AD NUMBER

B118 782

CLASSIFICATION CHANGES

TO

FROM

AUTHORITY

CNA ltr
15 Dec. 88

CRM 85-79 / August 1985

L
(2)

RESEARCH MEMORANDUM

ON THE CONTENT AND MEASUREMENT VALIDITY OF HANDS-ON JOB PERFORMANCE TESTS

Milton H. Maier
Catherine M. Hiatt

A Division of CNA Hudson Institute
CENTER·FOR·NAVAL·ANALYSES
4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268 • (703) 824-2000

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b RESTRICTIVE MARKINGS adm/oper use 9 mar 88
2a SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION / AVAILABILITY OF REPORT Distribution limited to U.S. Government agencies only; ~~Operational/ Administrative information contained~~. Other requests for this document must be referred to ~~the Deputy Chief of Staff (RD&S)~~.
2b DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) CRM 85-79	5 MONITORING ORGANIZATION REPORT NUMBER(S)

6a NAME OF PERFORMING ORGANIZATION Center for Naval Analyses	6b OFFICE SYMBOL (*If applicable*) CNA	7a NAME OF MONITORING ORGANIZATION Deputy Chief of Staff (RD&S)	
6c ADDRESS (*City, State, and ZIP Code*) 2000 North Beauregard Street Alexandria, Virginia 22311		7b ADDRESS (*City, State, and ZIP Code*) Headquarters, Marine Corps Washington, DC 20380	
8a NAME OF FUNDING / ORGANIZATION Office of Naval Research	8b OFFICE SYMBOL (*If applicable*) ONR	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N00014-83-C-0725	
8c ADDRESS (*City, State, and ZIP Code*) 800 North Quincy Street Arlington, Virginia 22217		10 SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO: 65154N PROJECT NO: R0148 TASK NO: WORK UNIT ACCESSION NO:	

11 TITLE (*Include Security Classification*)
On the Content and Measurement Validity of Hands-On Job Performance Tests

12 PERSONAL AUTHOR(S)
Milton H. Maier, Catherine M. Hiatt

13a TYPE OF REPORT Final	13b TIME COVERED FROM TO	14 DATE OF REPORT (*Year, Month, Day*) August 1985	15 PAGE COUNT 52

16 SUPPLEMENTARY NOTATION

17 COSATI CODES		
FIELD	GROUP	SUB-GROUP
05	09	

18 SUBJECT TERMS (*Continue on reverse if necessary and identify by block number*)
Aptitude tests, ASVAB (Armed Services Vocational Aptitude Battery), Enlisted personnel, Job performance, Marine Corps, Occupational specialties, Performance tests, Qualifications, Selection, Skills, Tables (Data)

19 ABSTRACT (*Continue on reverse if necessary and identify by block number*)

Hands-on tests of job performance have intrinsic validity because of their high fidelity to job behavior. However, they are susceptible to poor content and measurement validity. The purpose of this analysis is to examine the content and measurement validity of prototype hands-on tests developed for three Marine Corps specialties – Ground Radio Repair, Automotive Mechanic, and Infantry Rifleman.

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT ☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS	21 ABSTRACT SECURITY CLASSIFICATION Unclassified	
22a NAME OF RESPONSIBLE INDIVIDUAL Major Robinson	22b TELEPHONE (*Include Area Code*) (703) 824-2643	22c OFFICE SYMBOL RDS-40

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted

All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS Adm/Oper Use 9 Mar 88
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION / AVAILABILITY OF REPORT Distribution limited to U.S. Government agencies only; ~~Operational/~~ ~~Administrative information contained~~. Other requests for this document must be referred to ~~the Deputy Chief of Staff (RD&S)~~.
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) CRM 85-79	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Center for Naval Analyses	6b. OFFICE SYMBOL (If applicable) CNA
7a. NAME OF MONITORING ORGANIZATION Deputy Chief of Staff (RD&S)	
6c. ADDRESS (City, State, and ZIP Code) 2000 North Beauregard Street Alexandria, Virginia 22311	7b. ADDRESS (City, State, and ZIP Code) Headquarters, Marine Corps Washington, DC 20380
8a. NAME OF FUNDING / ORGANIZATION Office of Naval Research	8b. OFFICE SYMBOL (If applicable) ONR
9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N00014-83-C-0725	
8c. ADDRESS (City, State, and ZIP Code) 800 North Quincy Street Arlington, Virginia 22217	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO: 65154N; PROJECT NO: R0148; TASK NO:; WORK UNIT ACCESSION NO:

11. TITLE (Include Security Classification)

On the Content and Measurement Validity of Hands-On Job Performance Tests

12. PERSONAL AUTHOR(S)
Milton H. Maier, Catherine M. Hiatt

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Final	FROM TO	August 1985	52

16. SUPPLEMENTARY NOTATION

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP
05	09	

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Aptitude tests, ASVAB (Armed Services Vocational Aptitude Battery), Enlisted personnel, Job performance, Marine Corps, Occupational specialties, Performance tests, Qualifications, Selection, Skills, Tables (Data).

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Hands-on tests of job performance have intrinsic validity because of their high fidelity to job behavior. However, they are susceptible to poor content and measurement validity. The purpose of this analysis is to examine the content and measurement validity of prototype hands- on tests developed for three Marine Corps specialties – Ground Radio Repair, Automotive Mechanic, and Infantry Rifleman.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS	Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE (Include Area Code)	22c. OFFICE SYMBOL
Major Robinson	(703) 824-2643	RDS-40

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted

All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE

A Division of CNA Hudson Institute

CENTER·FOR·NAVAL·ANALYSES

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

4 November 1985

MEMORANDUM FOR THE DISTRIBUTION LIST

Subj: Center for Naval Analyses Research Memorandum 85-79

Encl: (1) CRM 85-79, "On the Content and Measurement Validity of Hands-on Job Performance Tests", by Milton H. Maier and Catherine M. Hiatt, August 1985

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

2. This Research Memorandum examines the content and measurement validity of hands-on job performance tests for two Marine Corps occupational specialties, Ground Radio Repair and Automotive Mechanic. The results point to the need for designing the tests to incorporate the full range of job requirements and exercising rigorous quality control over test administrations.

Christopher Jehn

Christopher Jehn
Director
Marine Corps Operations
Analysis Group

Distribution List:
Reverse page

Subj: Center for Naval Analyses Research Memorandum 85-79

Distribution List

Job Performance Measurement Working Group
(Attn: LTC. Harris, Directorate of Accession Policy/OSD) (18 copies)
Selection and Classification Working Group
(Attn: Dr. Malcolm Ree, AFHRL/MOAE
Dr. Clarence McCormick, Hq MEPCOM/MEPCT-P
Mr. Paul Foley, NPRDC, Code 310)
National Academy of Sciences Advisory Committee
(Attn: Ms. Wigdor, National Research Council/NAS) (20 copies)
Navy Personnel Research and Development Center
Chief of Naval Education and Training
Chief of Naval Technical Training
Army Research Institute
Air Force Human Resources Laboratory
Deputy Chief of Staff for Manpower (Attn: MPI-20) (2 copies)
Deputy Chief of Staff for Research, Development and Studies (2 copies)

CRM 85-79 / August 1985

ON THE CONTENT AND MEASUREMENT VALIDITY OF HANDS-ON JOB PERFORMANCE TESTS

Milton H. Maier
Catherine M. Hiatt

Marine Corps Operations Analysis Group

A Division of CNA *Hudson Institute*

CENTER·FOR·NAVAL·ANALYSES

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

ABSTRACT

Hands-on tests of job performance have intrinsic validity because of their high fidelity to job behavior. However, they are susceptible to poor content and measurement validity. The purpose of this analysis is to examine the content and measurement validity of prototype hands-on tests developed for three Marine Corps specialties – Ground Radio Repair, Automotive Mechanic, and Infantry Rifleman.

EXECUTIVE SUMMARY

PROBLEM

The justification for using aptitude tests to help select enlisted recruits and assign them to occupational specialties is that aptitude tests are valid predictors of performance. The aptitude tests used by the military services have been extensively validated as predictors of performance in occupational specialty training courses. Their usefulness as predictors of performance on the job, however, is less well documented. The Job Performance Measurement Project has been initiated to validate the Armed Services Vocational Aptitude Battery (ASVAB) as a predictor of job performance.

The question then arises of how job performance should be measured. The measures favored by the Joint Service Job Performance Measurement Working Group are hands-on job performance tests. These tests have intrinsic validity because of their high fidelity to job behavior. Hands-on performance tests, however, are susceptible to poor content and measurement validity.

- Poor content validity may arise because the tests focus on skills easy to test in the hands-on mode without including the full range of job requirements.

- Poor measurement validity may arise because the scoring standards of test administrators are not calibrated and because test security is difficult to maintain (examinees can find out what is being tested and practice beforehand).

The purpose of this analysis is to examine the content and measurement validity of prototype hands-on tests for three Marine Corps specialties—Ground Radio Repair, Automotive Mechanic, and Infantry Rifleman—used in a feasibility study to evaluate ASVAB qualification standards.

FINDINGS

The findings pertain to the two technical specialties, Ground Radio Repair and Automotive Mechanic. Because the infantry riflemen in the sample had limited job experience, the results for them are inconclusive.

- Hands-on test scores were only weakly related to amount of job experience, as measured by months in the Marine Corps (figure I). Test scores were expected to increase with experience, and the lack of relationship raises questions about how well the tests represent the full range of job requirements.

- The ASVAB is a valid predictor of hands-on test scores for people with 2 years or less of service in the Marine Corps, but not for people with more than 2 years of service (table I).

- Hands-on test scores did increase with experience for people with low aptitude, but not for people with high aptitude (figure II).

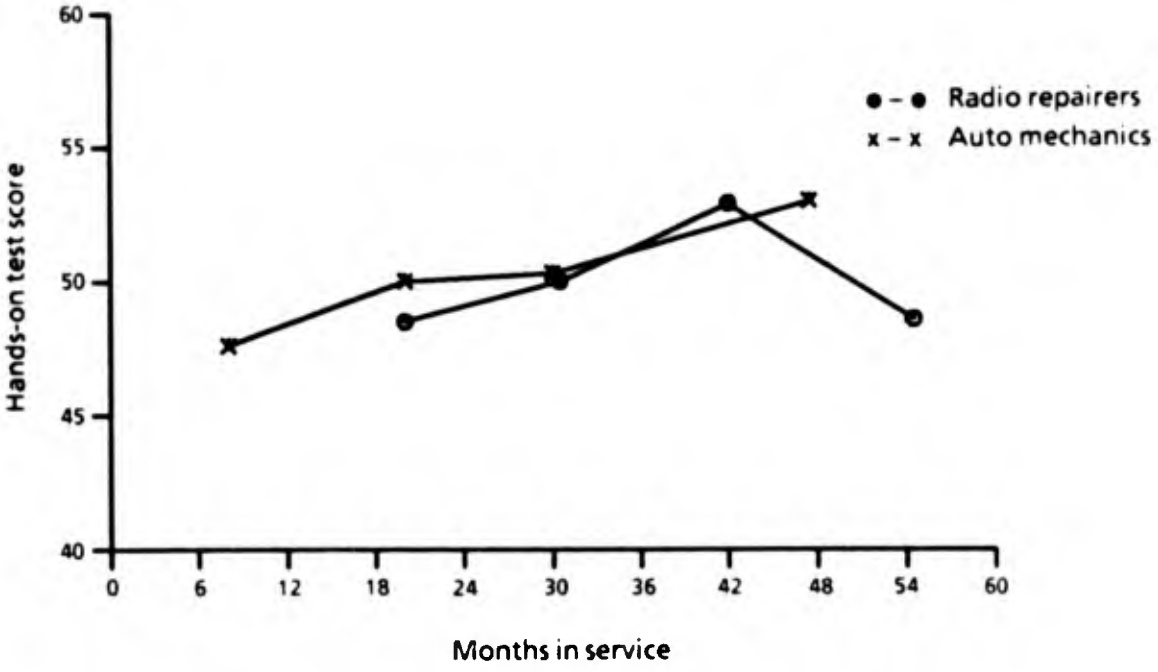


FIG. I: JOB PERFORMANCE RELATED TO TIME IN THE MARINE CORPS

These results suggest that the hands-on test content was appropriate for people recently assigned to their first duty station, but less appropriate for people with more experience, who perform job tasks not reflected in the tests.

The findings on measurement validity bear on institutionalizing of hands-on job performance tests:

- The test administrators used different scoring standards, and the same administrators changed their scoring standards across time (see figure III).

- Maintaining test security is difficult.

TABLE I

VALIDITY OF THE ASVAB FOR PREDICTING HANDS-ON TEST SCORES

Months in service	Validity[a]	Number of cases
	Ground Radio Repair	
15-25	.69	38
26-35	.00	53
36-48	.00	37
Total	.37	128
	Automotive Mechanic	
2-14	.72	57
15-25	.52	56
26-34	.15	53
35-60	-.07	54
Total	.37	220

a. Population-wide estimate of validity coefficient.

CONCLUSIONS

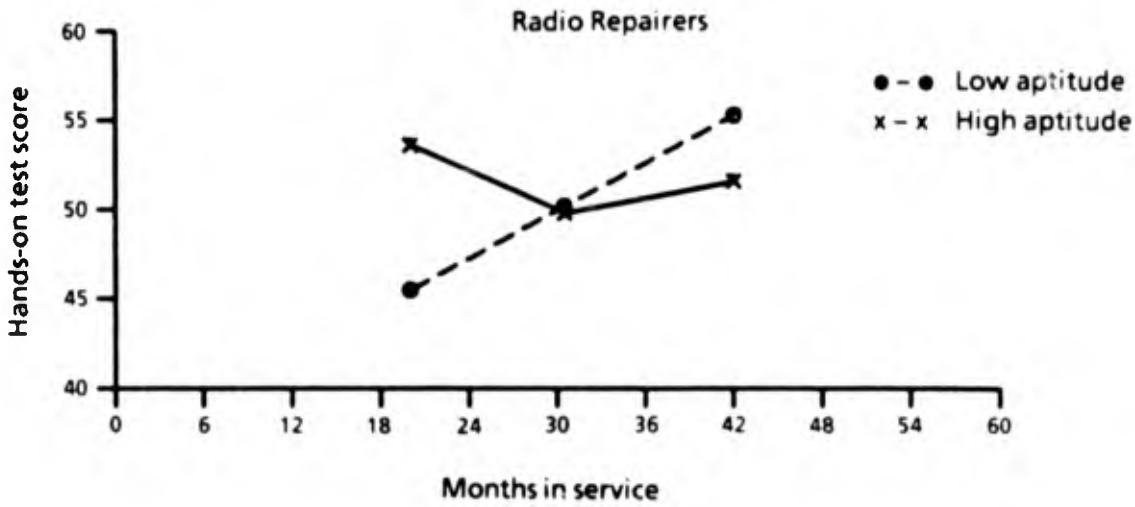
- The ASVAB is a valid predictor of job performance, as measured by hands-on tests.

- But hands-on tests lack robustness:

  - Content validity is sensitive to job experience.

  - Measurement validity is sensitive to the calibration, or scoring standards, of test administrators.

- Institutionalizing hands-on job performance tests would be difficult.



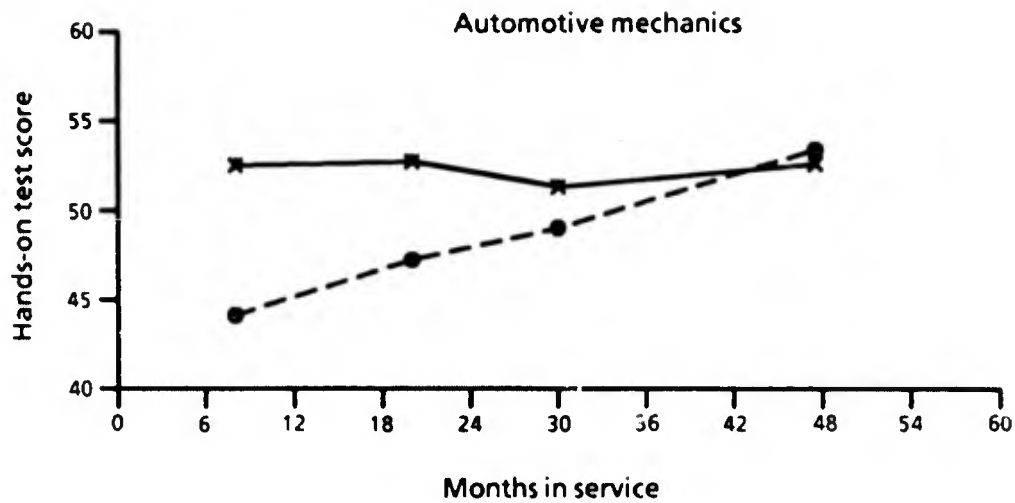


FIG. II: PERFORMANCE RELATED TO APTITUDE AND EXPERIENCE

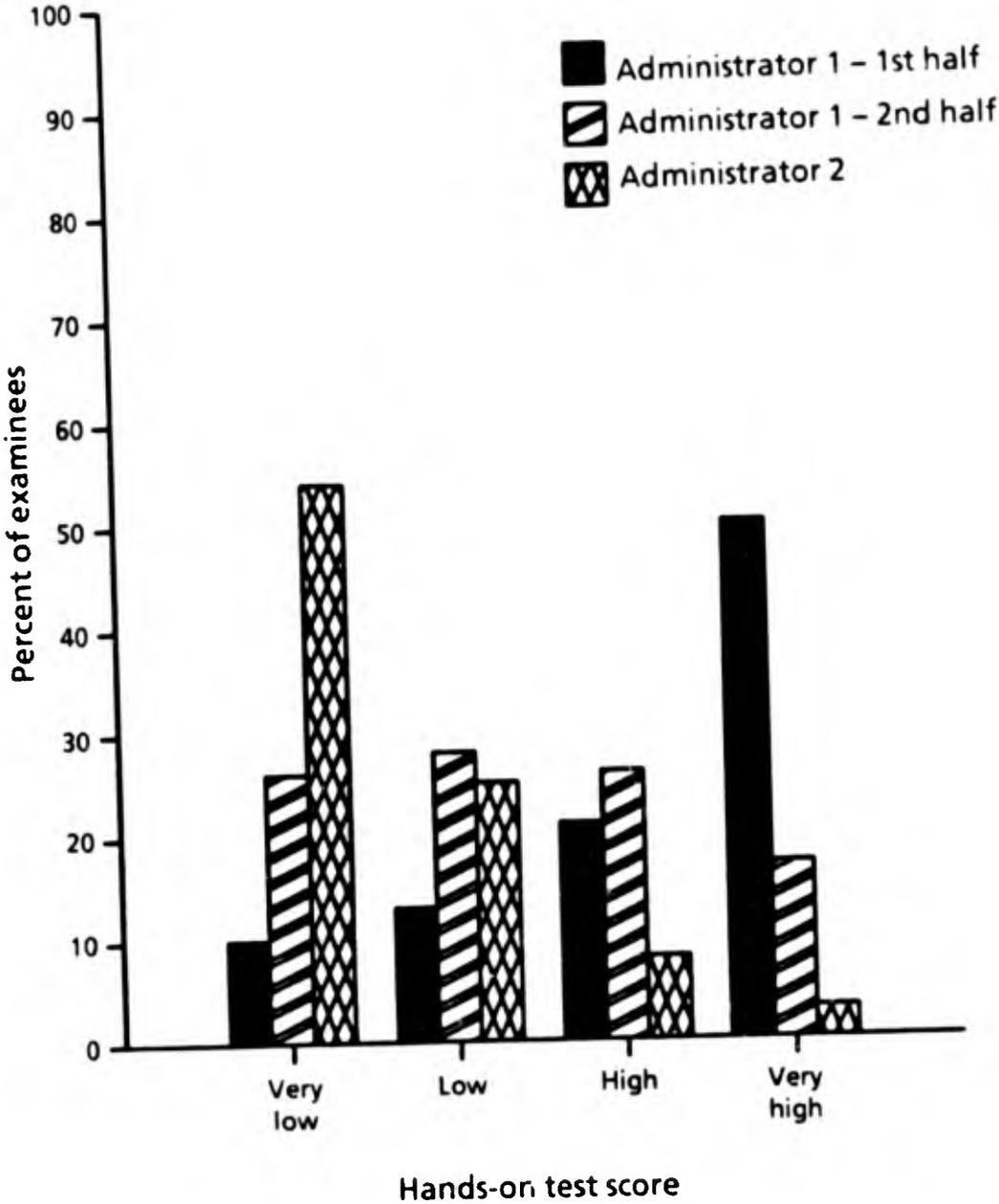


FIG. III: HANDS-ON TEST SCORES ASSIGNED BY TEST ADMINISTRATORS TO AUTOMOTIVE MECHANICS

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

SECTION 1

INTRODUCTION

Hands-on job performance tests have intrinsic validity because of their high fidelity to the skills required to perform job tasks. For this reason, the Joint Service Job Performance Measurement Working Group (JPMWG)[1] has decided that hands-on tests are the benchmark measure for evaluating the job proficiency of people in their occupational specialties and for evaluating the job relatedness of surrogate measures of job performance, such as written tests, ratings, and grades in occupational specialty training courses. Hands-on tests are the key to validating enlistment standards and to finding surrogate measures of job performance that are feasible for routine use in the military personnel systems.

THREATS TO THE VALIDITY OF HANDS-ON TESTS

Despite their high fidelity to job skills, certain characteristics inherent in the nature of hands-on tests can seriously threaten their validity as measures of job performance. Because of the difficulty of incorporating the full range of job requirements into hands-on tests, these tests run a high risk of being focused too narrowly on those job requirements that are readily amenable to testing in the hands-on mode. Hands-on tests are resource intensive – that is, they require expensive test administrators, frequently tie up expensive equipment, and take a long time to administer. Typically, 15 to 30 tasks are covered by a hands-on test, and the time limit imposed by the military services is about 4 hours. Because of the constraints of time and feasibility, ordinarily only a portion of the full range of job requirements can be included in a hands-on test.

The job requirements most likely to be left out of hands-on tests involve complex skills; examples are preparing written documents; operating complex equipment, such as driving a tank; or expending costly hardware, such as firing a missile. The content of hands-on tests frequently is limited to the

1. The JPMWG is composed of technical and policy representatives from each service and is chaired by the Office of the Secretary of Defense.

relatively simple job requirements that can be easily observed and evaluated, take only a short time to accomplish, and do not consume or are not likely to destroy expensive equipment. To the extent that hands-on tests focus only on job requirements that are easy to test, their content validity is degraded.

The measurement validity of hands-on tests may be lowered by the way they are administered and scored. Typically, each administrator tests a single examinee at a time and scores the examinee's performance while administering the test by judging the correctness of the examinee's responses. Because of the ambiguity inherent in evaluating human behavior, administrators have latitude in assigning scores to examinees.

In hands-on testing, administrators set the standard for evaluating performance. To obtain valid test scores, the scoring standards of test administrators must be accurately calibrated. At a minimum, different administrators observing the test performance of the same examinee should agree on how they score the test. In addition, the scores should accurately reflect how well the examinee's responses satisfy the job requirements incorporated in the test. To the extent that administrators' scoring standards disagree or shift, either randomly or in a constant direction, their calibration is inaccurate and measurement validity is lowered.

Measurement validity may also be lowered through faulty test security. Hands-on testing ordinarily extends over time, and examinees tested later in a given period can learn the test content from those tested earlier. Through practice, some examinees can learn to perform the specific tasks in the test, but still perform poorly overall in the specialty.

These threats to the content and measurement validity of hands-on tests can be controlled through careful testing procedures. Job requirements can be adequately represented if test content is judiciously selected. Consistent and accurate scoring can be maintained by carefully training and monitoring test administrators. Test security can be enhanced by admonishing examinees to refrain from discussing the tests with other potential examinees or by developing multiple forms of hands-on tests for each specialty. A major goal of the Job Performance Measurement Project is to develop effective procedures for constructing and administering hands-on tests. The decision to use hands-on tests as the benchmark measure indicates the confidence of the JPMWG that this goal can be accomplished.

EFFECTS ON THE VALIDITY OF QUALIFICATION STANDARDS

Qualification standards for assigning enlisted recruits to occupational specialties are based primarily on the Armed Services Vocational Aptitude Battery (ASVAB). The ASVAB measures aptitudes in a paper-and-pencil format. The justification for using the ASVAB to set qualification standards is that people with qualifying ASVAB scores have a higher probability of being successful performers in the occupational specialty training courses than people with failing ASVAB scores. Therefore, the validity of the tests used to measure job performance has serious consequences for validating qualification standards.

To the extent that hands-on test scores do not reflect realistic job requirements, the predictive validity of the ASVAB is likely to be lowered. In other words, low content validity will probably lower the predictive validity of the ASVAB. A lack of measurement validity, arising from poor calibration of test administrators or from poor test security, is certain to lower the predictive validity of the ASVAB.

Based on the extensive validation studies conducted by the military services, the ASVAB should be a reasonably accurate predictor of job performance in the normal job environment, where people have an opportunity to learn how to perform job tasks and can exercise their skills on the job. However, for people who have not learned the task or have not performed it for a long period, while others perform it frequently, the ASVAB will appear to be a poor predictor of performance. Thus, job experience may be a powerful variable affecting the relationship between ASVAB and hands-on test scores.

The Marine Corps conducted a study to evaluate the feasibility of using hands-on job performance tests for three occupational specialties to validate ASVAB qualification standards [1]. The results of the study were highly favorable; the appropriate ASVAB aptitude composites had respectable validity coefficients (.55 to .60). The qualification standards derived in the feasibility study agreed closely with the existing standards used by the Marine Corps. In the study, the effects of measurement error were partially controlled by statistically removing differences among scoring standards of test administrators; that is, the hands-on test scores for each administrator were standardized to have the same mean and standard deviation. No thorough analysis of the content validity of the hands-on tests was made in the feasibility study.

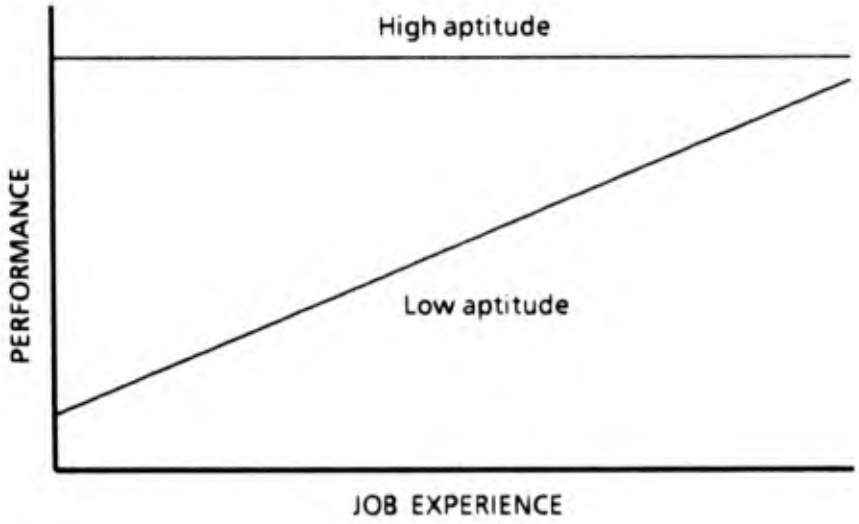
PURPOSES OF THIS ANALYSIS

The purposes of this analysis are the following:

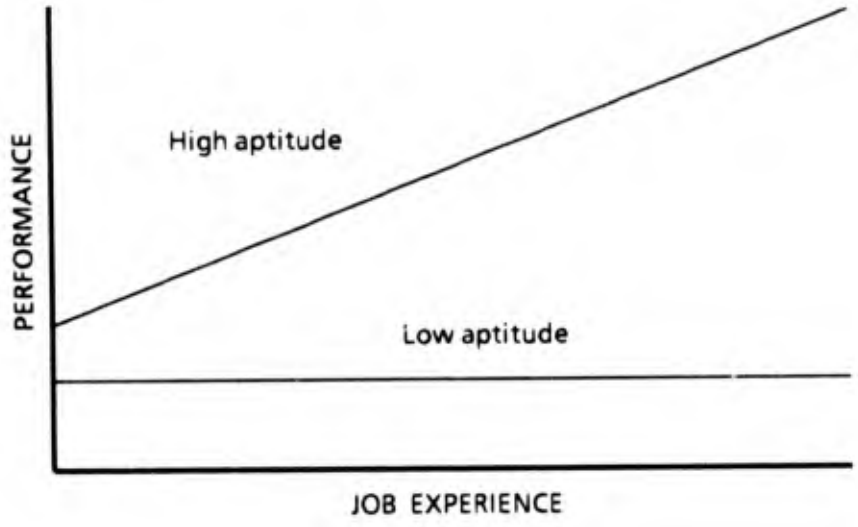
- To evaluate the content validity of the hands-on tests used in the Marine Corps feasibility study.
- To evaluate the effects of poor measurement validity on hands-on test scores.
- To consider the implications of large-scale, routine use of hands-on testing in the military.

Three hypotheses are advanced about how content validity can be evaluated by examining the way in which job experience and aptitude are related to hands-on test performance.

- If the content of the hands-on tests focuses on minimal skills and knowledge required of entry-level people with limited experience, people who are proficient on the more complex job requirements could not demonstrate the breadth and depth of their competence. As a rule, people with high aptitude would quickly become proficient on the test content and then show little gain with increased job experience, whereas people with low aptitude would start with a lower proficiency but, with experience, catch up to those with high aptitude. Portrayed graphically, the joint relationship of aptitude and experience to job performance would look like this:

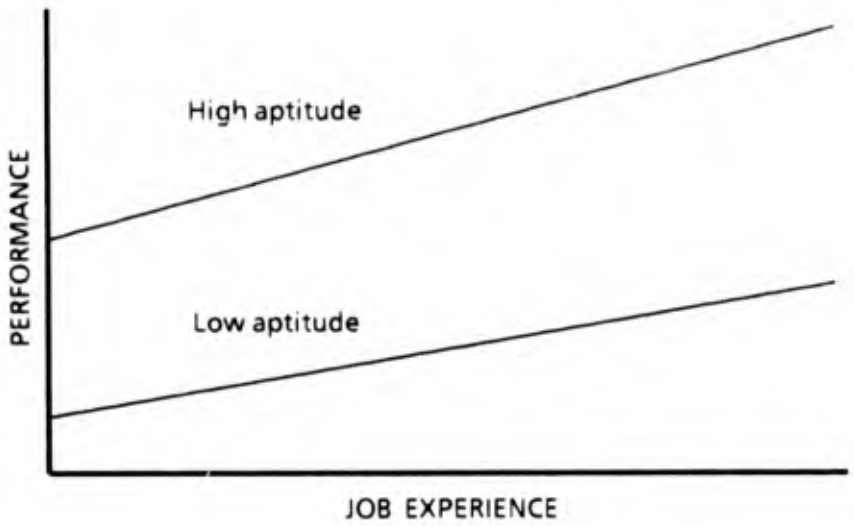


- If the hands-on test focuses on complex job requirements that normally are performed only by the more competent workers, people with high aptitude would tend to be initially more proficient than people with low aptitude and the gap between the two groups would grow as experience increases; that is, on complex tasks, people with high aptitude profit more from experience than people with low aptitude. This reasoning is based on the widespread practices, among employers, of requiring high aptitude of managers and professionals and, among colleges, of using aptitude scores to select students. The military services require commissioned officers to have aptitude scores about one standard deviation above the mean. If the hands-on test focuses on complex job requirements, the joint relationship of aptitude and experience to job performance would look like this:



- If the test content reflects the full range of job requirements, test scores should increase as both job experience and aptitude increase. Assuming that, over the years, the Marine Corps has learned to structure job requirements to match the aptitudes of people typically assigned to the various specialties, people with more experience should be proficient on more tasks. Given that the ASVAB is an established predictor of performance in occupational specialty training courses, it can reasonably be expected to have a significant relationship to the breadth and depth of the skills and knowledge underlying the full range of job requirements. If the hands-on test

has adequate content validity, the joint relationship of aptitude and job experience to job performance should look like this:



All three hypotheses indicate that performance should increase as experience increases. This expectation pervades military and civilian personnel systems. Pay scales and promotion are tied directly to length of time served in the military. Seniority is an important factor in most civilian pay scales. This notion implies that if hands-on tests have content validity, scores should increase with job experience.

PROCEDURES

The data for this analysis originally were collected as part of the Marine Corps study to evaluate the feasibility of using hands-on job performance tests to validate ASVAB qualification standards [1]. The three specialties evaluated were Ground Radio Repair, with high technical job requirements (37 weeks of formal school training); Automotive Mechanics, with moderate technical job requirements (13 weeks of training); and Infantry Rifleman, with low technical job requirements (5 weeks of training). Marine Corps job experts, assisted by testing psychologists, developed both a hands-on test and a written test for each specialty [2]. The tests were administered by other Marine Corps job experts detailed to the project. They were not specially trained on how to administer or score the tests, and no quality control measures were exercised over the testing.

The sample sizes were:

Ground Radio Repair —137 cases
Automotive Mechanic —220 cases
Infantry Rifleman —161 cases

Most examinees were in their first term of enlistment; a few from the two technical specialties (Radio Repair and Automotive Mechanic) were in their second term. All examinees in this analysis had taken forms 6 or 7 of the ASVAB (ASVAB 6/7), which were used from January 1976 until October 1980. The job performance testing started in mid-July 1981 and was finished in late 1981.

All testing was done at Camp Pendleton, California. The examinees were selected by local units. Virtually all Marine Corps ground radio repairers and a large percentage of the auto mechanics in southern California were tested. The riflemen, however, were only a small proportion of the available examinees; they tended to be recent graduates from the Infantry Training School, with little experience in their duty assignment.

The content validity of the hands-on tests was evaluated by determining how the scores related to other variables commonly viewed as indicators of performance. The first check was on the relationship between hands-on test scores and job experience, measured by months of service in the Marine Corps. Other variables used to check content validity were written proficiency tests, grades in occupational specialty training courses, proficiency ratings by supervisors, irregular discharges, reductions in grade, and promotion rates.

Measurement validity was evaluated by examining how different scoring standards among test administrators affected the examinees' scores in the two technical specialties. In these specialties, a single administrator gave the entire hands-on test to an examinee. The hands-on test scores, therefore, were systematically affected by a lack of calibrating the administrators' scoring standards. In the Rifleman specialty, the content of the hands-on test was divided into several parts, with each part administered at a separate testing station. A test administrator was assigned to each station, and examinees moved from station to station. Because each administrator tested virtually all examinees and each examinee was tested by virtually all administrators, any consistent differences in scoring standards among the adminstrators affected all examinees in the same way.

Measurement validity was also evaluated by determining how test scores were related to the time the testing took place. The examinees were grouped by testing date, and changes in the distributions of test scores were evaluated.

In the analysis of content validity, the hands-on test scores assigned by each administrator in the technical specialties were standardized to have the same mean and standard deviation. This standardization helped remove differences in scoring standards among the administrators but did not affect any differences in the accuracy of observing and scoring test performance. Note that the analysis of measurement validity used the hands-on test scores as assigned by the administrator.

SECTION 2

RESULTS

The results bearing on the content validity and the measurement validity of the hands-on tests are presented below. The written test is also included as a measure of performance for comparison of content validity. Because the measurement validity of the written tests is not an issue, no results for them are included in the second subsection.

CONTENT VALIDITY OF HANDS-ON TESTS

The first analysis examined the relationship between hands-on test scores and job experience, measured as months in the Marine Corps. The samples of examinees in each specialty were grouped by level of experience, and mean performance test scores were computed for each level. The mean scores are plotted in figure 1. The top panel of figure 1 shows the mean hands-on test scores, which are standardized to have a mean of 50 and a standard deviation of 10. As noted earlier, the hands-on test scores assigned by each test administrator in the two technical specialties had already been standardized to have a mean of 50 and standard deviation of 10, which helped remove differences in calibration among the administrators. The middle panel shows the mean efficiency scores, which are the standardized hands-on test scores divided by testing time in minutes. Efficiency scores were also standardized to have a mean of 50 and a standard deviation of 10 in each sample. No efficiency scores were computed for the sample of riflemen because testing time was not a meaningful variable for their hands-on test. The bottom panel shows the mean scores for the written tests, standardized to a mean of 50 and standard deviation of 10. For each specialty, the mean performance scores are shown for the mid-point of the experience intervals (which are listed in tables 1, 2, and 3).

The plots of the mean hands-on test score and efficiency score are essentially flat for each specialty. Contrary to expectations, hands-on test scores did not increase with job experience. However, the mean scores on the written tests did show an orderly increase with experience. The statistical significance of difference among the means was computed by analysis of variance. The results are shown in table 1 for radio repairers, table 2 for

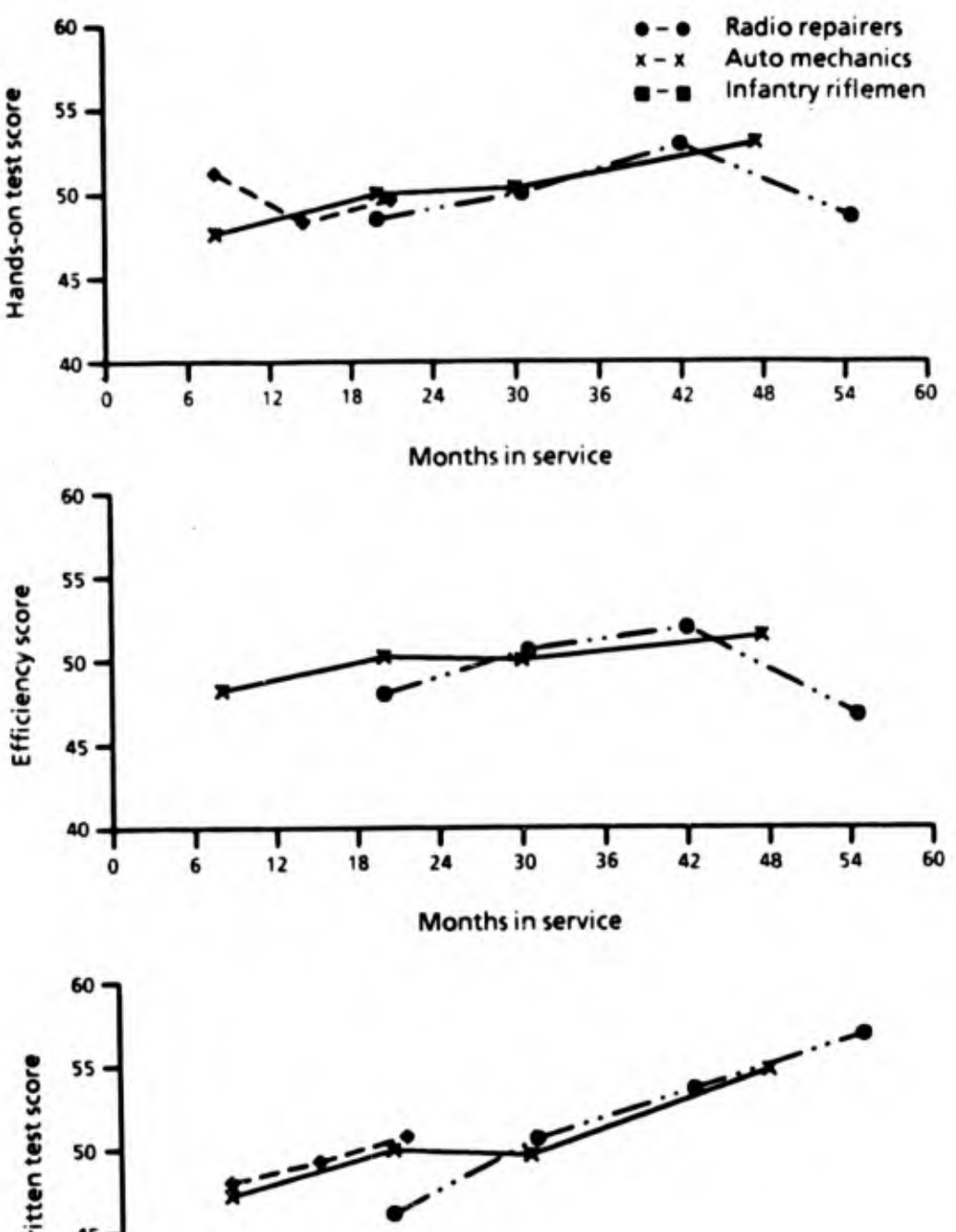


FIG. 1: MEAN JOB PERFORMANCE TEST SCORES RELATED TO MONTHS IN THE MARINE CORPS

TABLE 1

ANALYSIS OF PERFORMANCE TEST SCORES BY LEVEL OF JOB EXPERIENCE – GROUND RADIO REPAIRERS

Part A: Levels of job experience

Months in Marine Corps	Number of cases
15-25	38
26-35	53
36-48	37
49-60	9
Total	137

Part B: Analysis of variance

Source of variation	Sum of squares	Degrees of freedom	Mean square	*F* value[a]
Hands-on test scores				
Between groups[b]	398	3	132.7	1.33
Within groups[c]	13,283	133	99.9	
Efficiency scores				
Between groups	406	3	135.3	1.36
Within groups	13,193	133	99.2	
Written test scores				
Between groups	1,381	3	460.3	4.86**
Within groups	12,602	133	94.7	

a. Mean square between groups divided by mean square within groups. *F* values significant at the 1-percent level are shown by **; those at the 5-percent level are shown by *

b. The between-groups variation is the index of how much job performance varies with levels of experience

c. The within-groups variation is the error term.

TABLE 2

ANALYSIS OF PERFORMANCE TEST SCORES BY LEVEL OF JOB EXPERIENCE – AUTOMOTIVE MECHANICS

Part A: Levels of job experience

Months in Marine Corps	Number of cases
2-14	57
15-25	56
26-34	53
35-60	54
Total	220

Part B: Analysis of variance

Source of variation	Sum of squares	Degrees of freedom	Mean square	*F* value[a]
Hands-on test scores				
Between groups[b]	823	3	274.2	2.84*
Within groups[c]	20,858	216	96.5	
Efficiency scores				
Between groups	423	3	141.2	1.41
Within groups	21,576	216	99.9	
Written test scores				
Between groups	1,493	3	497.7	5.65**
Within groups	19,027	216	88.1	

a. Mean square between groups divided by mean square within groups. *F* values significant at the 1-percent level are shown by **; those at the 5-percent level shown are by *.
b. The between-groups variation is the index of how much job performance varies with levels of experience.
c. The within-groups variation is the error term.

TABLE 3

ANALYSIS OF PERFORMANCE TEST SCORES BY LEVEL OF JOB EXPERIENCE – INFANTRY RIFLEMEN

Part A: Levels of job experience

Months in Marine Corps	Number of cases
3-13	42
14-15	47
16-26	63
Total	152

Part B: Analysis of variance

Source of variation	Sum of squares	Degrees of freedom	Mean square	*F* value[a]
Hands-on test scores				
Between group[b]	181	2	90.3	0.94
Within group[c]	14,321	149	96.1	
Written test scores				
Between groups	183	2	91.7	1.17
Within groups	11,697	149	78.5	

a. Mean square between groups divided by mean square within groups. *F* values significant at the 1-percent level are shown by **; those at the 5-percent level are shown by *

b. The within-groups variation is the error term.

c. The between-groups variation is the index of how much job performance varies with levels of experience.

mechanics, and table 3 for riflemen. The mean performance scores for each level of experience are shown in appendix A.

The relationship between hands-on test scores and job experience is not statistically significant, except in the sample of mechanics. In this case, the total hands-on test scores, but not the efficiency scores, are marginally significant (at the 5-percent level). The written tests are significantly related to job experience (at the 1-percent level) for the samples of radio repairers and mechanics, but not for the riflemen.

The general lack of relationship between hands-on test scores and job experience raises questions about the content validity of the hands-on tests. Three hypotheses were advanced in the Introduction on how the interaction between aptitude and job experience could help clarify the content validity of the hands-on tests. Given that experience is only weakly related to the hands-on tests used in the feasibility study, the third hypothesis, that performance increases with both experience and aptitude, is, of course, not tenable. Further analyses were conducted to evaluate hypotheses 1 and 2, which discuss how content validity can be inferred from the effect of aptitude and experience on job performance.

Interaction Effects of Aptitude and Job Experience

Each sample was divided into high and low levels of aptitude at the median of the appropriate aptitude composite (Electronics Repair for radio repairers, Mechanical Maintenance for automotive mechanics, and General Technical for infantry riflemen[1]). The intervals of job experience were the same as those used in the previous analysis. The interaction effects of aptitude and experience were computed by analysis of variance, except that

1. In this analysis the aptitude composites were defined as they are for forms 11, 12, and 13 of the ASVAB. Electronics Repair contains the General Science, Arithmetic Reasoning, Math Knowledge, and Electronics Information subtests; Mechanical Maintenance contains the Arithmetic Reasoning, Auto/Shop Information (the Auto Information subtest from ASVAB 6/7 was used), Mechanical Comprehension, and Electronics Information subtests; General Technical contains the Verbal, Arithmetic Reasoning, and Mechanical Comprehension subtests. The current definitions of aptitude composites were used in this analysis because they are expected to generalize better to current and future qualification standards.

the nine second-term Marines (with over 48 months of experience) were removed from the Radio Repair sample.

The mean performance test scores for each combination of aptitude score and job experience are shown in figure 2 for radio repairers, in figure 3 for automotive mechanics, and in figure 4 for infantry riflemen. The same performance scores were used as in the previous analysis (total hands-on test score, efficiency score, and written test scores). The results of the analysis of variance are shown in tables 4, 5, and 6 for the three specialties. The means and standard deviations are provided in appendix A.

In the two technical specialties (figures 2 and 3), people with high aptitude had high hands-on and efficiency scores at low levels of experience, but their scores did not increase as experience increased. For people with low aptitude, hands-on test scores did increase as their experience increased, to levels equal to or higher than those for people with high aptitude. Scores on the written test tended to increase uniformly with experience for both levels of aptitude.

The results for the infantry riflemen differed from those for the technical specialties. Riflemen with high aptitude consistently scored higher than those with low aptitude on both the hands-on and written tests. The experience levels, however, were relatively low (2 years or less in the Marine Corps), and there was an indication that at the end of 2 years, people with low aptitude were catching up on both the hands-on and written tests.

The statistical significance of the relationships conform to the pattern of means. The interaction between aptitude and experience was significant at the 5-percent level in the radio repair sample when the hands-on score was the dependent variable (table 4). When the written test was the dependent variable, the interaction effects were not significant, but aptitude was highly related to the written test scores and experience was marginally related (F was significant at 5.6-percent level).

In the sample of automotive mechanics (table 5), the interaction effects were significant (at the 1-percent level) when efficiency scores were the dependent variable, but not significant when hands-on test scores were the dependent variable. Aptitude was significantly related to all three performance measures (at the 1-percent level). Experience was related to the written test scores, but not to the hands-on or efficiency scores.

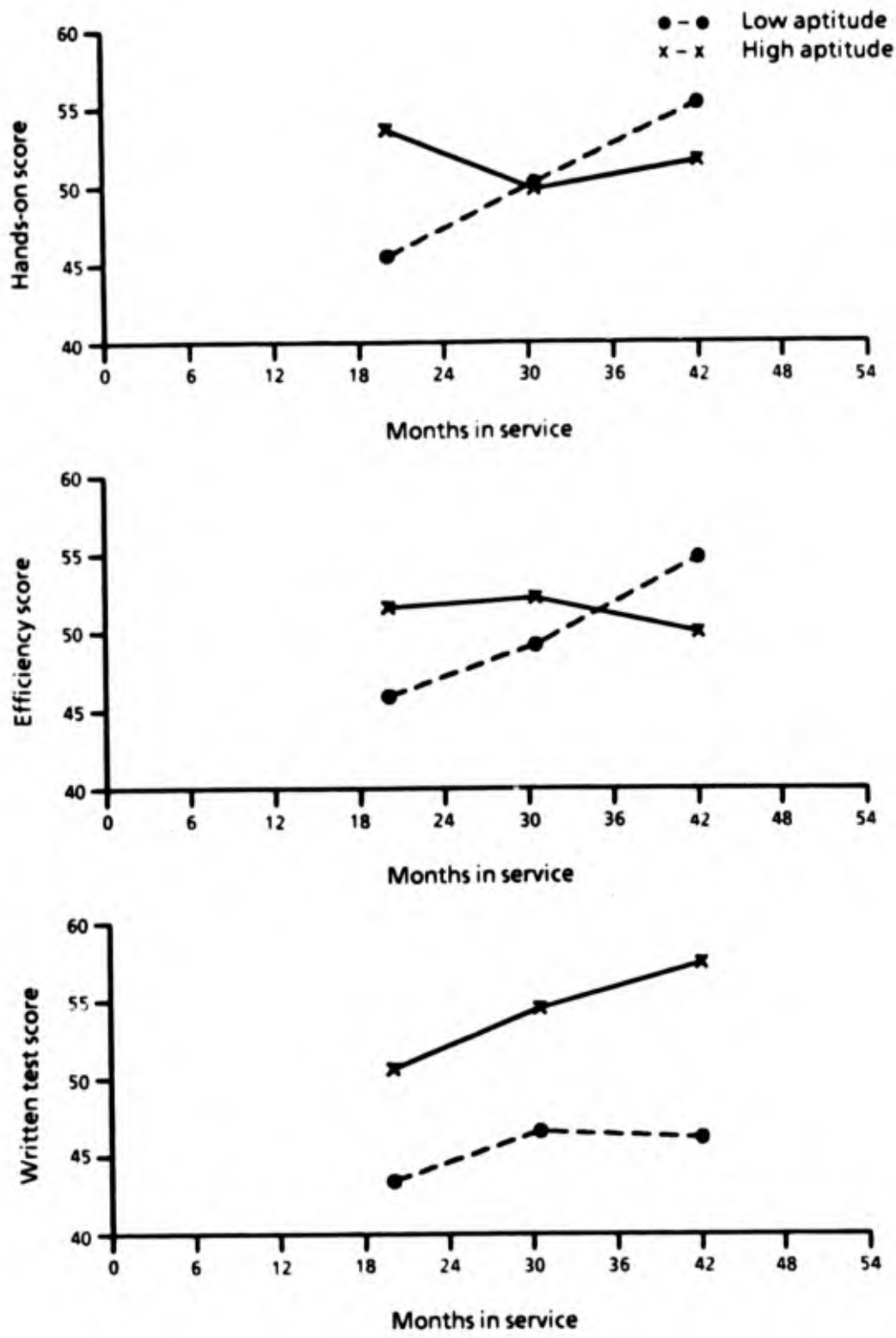


FIG. 2: JOB PERFORMANCE RELATED TO APTITUDE AND EXPERIENCE – GROUND RADIO REPAIRERS

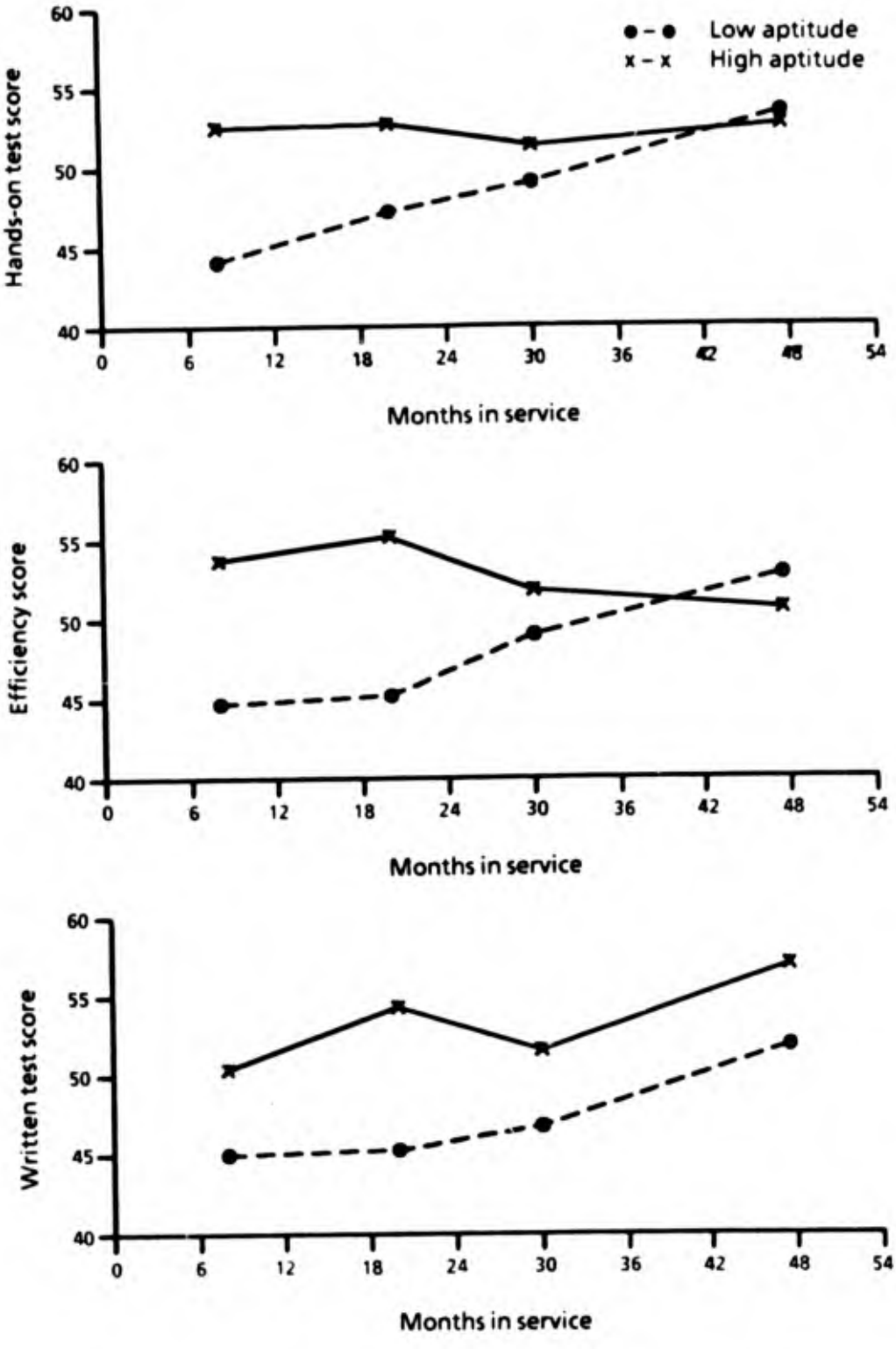


FIG. 3: JOB PERFORMANCE RELATED TO APTITUDE AND EXPERIENCE – AUTOMOTIVE MECHANICS

FIG. 4: JOB PERFORMANCE RELATED TO APTITUDE AND EXPERIENCE – INFANTRY RIFLEMEN

TABLE 4

EFFECTS OF APTITUDE AND JOB EXPERIENCE
ON PERFORMANCE MEASURES – GROUND RADIO REPAIRERS

Source of variation	Sum of squares	Degrees of freedom	Mean square	*F* value[a]
Hands-on test				
Aptitude	40	1	40	0.4
Experience	309	2	155	1.6
Interaction (Apt x Exp)	659	2	330	3.4*
Error[b]	11,944	122	98	
Efficiency score				
Aptitude	197	1	197	0.9
Experience	478	2	239	1.1
Interaction (Apt x Exp)	1,023	2	511	2.4
Error	25,838	122	212	
Written test				
Aptitude	2,266	1	2,266	28.9**
Experience	461	2	230	2.9
Interaction (Apt x Exp)	80	2	40	0.5
Error	9,553	122	78	

a. Mean square between groups divided by mean square within groups. *F* values significant at the 1-percent level are shown by **; those at the 5-percent level shown by *

b. The within-groups variation is the error term

TABLE 5

EFFECTS OF APTITUDE AND JOB EXPERIENCE ON PERFORMANCE MEASURES – AUTOMOTIVE MECHANICS

Source of variation	Sum of squares	Degrees of freedom	Mean square	*F* value[a]
Hands-on test				
Aptitude	806	1	806	8.8**
Experience	690	3	230	2.5
Interaction (Apt × Exp)	639	3	213	2.3
Error[b]	19,417	212	92	
Efficiency score				
Aptitude	1,344	1	1,344	14.6**
Experience	208	3	69	0.8
Interaction (Apt × Exp)	1,333	3	444	4.8**
Error	19,533	212	92	
Written test				
Aptitude	2,022	1	2,022	25.6**
Experience	1,370	3	457	5.8**
Interaction (Apt × Exp)	166	3	55	0.7
Error	16,762	212	79	

a. Mean square between groups divided by mean square within groups. *F* values significant at the 1-percent level are shown by **; those at the 5-percent level are shown by *.

b. The within-groups variation is the error term.

TABLE 6

EFFECTS OF APTITUDE AND EXPERIENCE
ON PERFORMANCE MEASURES – INFANTRY RIFLEMEN

Source of variation	Sum of squares	Degrees of freedom	Mean square	*F* value[a]
Hands-on scores				
Aptitude	1,234	1	1,234	14.0**
Experience	228	2	114	1.3
Interaction (Apt x Exp)	189	2	95	1.1
Error[b]	12,906	146	88	
Written test				
Aptitude	1,680	1	1,680	25.5**
Experience	183	2	92	1.4
Interaction (Apt x Exp)	397	2	199	3.0*
Error	9,621	146	66	

a. Mean square between groups divided by mean square within groups. *F* values significant at the 1-percent level are shown by **; those at the 5-percent level are shown by *.
b. The within-groups variation is the error term.

The patterns of mean differences for the two technical specialties support the hypothesis that the test content is appropriate for people with limited job experience, but less so for people with more experience and high aptitude. The fact that hands-on performance scores did not change for people with high aptitude, whereas written test scores increased, suggests that the hands-on tests did not permit the more able examinees to demonstrate the full extent of their job proficiency. The range of experience in the sample of riflemen was too limited for a good check on the interaction effects, but the data suggest that people with low aptitude improved their hands-on test scores, whereas people with high aptitude did not.

The presence of a significant interaction between aptitude and experience has strong implications for the predictive validity of the ASVAB and, hence, for the validity of qualification standards. As noted earlier, the

ASVAB-based qualification standards are justified only if the ASVAB accurately predicts performance. If people with low aptitude can compensate for their slowness in learning through job experience and learn to perform as well as or better than people with high aptitude, then the current practice of using the ASVAB to select recruits must be reevaluated. If, on the other hand, the test content is not appropriate for people with high aptitude and job experience, then the qualification standards may be valid, but the content of the hands-on tests must be made to better reflect actual job requirements.

Validity of the ASVAB by Level of Experience

As a further check on the predictive validity of the ASVAB, the correlation of aptitude scores with performance scores was computed for different levels of experience. The validity of the appropriate ASVAB aptitude composite for predicting hands-on test scores in each specialty is shown in table 7. The examinees in each sample were grouped by months in the Marine Corps, and the validity computed for each subgroup. The validity coefficients in each subgroup were corrected for range restriction, using the univariate model, to obtain population-wide estimates that put all coefficients on a common metric. The standard deviation of each aptitude composite in the population is 20. The standard deviation of the aptitude composite scores and the number of cases in each subgroup and the total sample are also shown in table 7.

The validity of the ASVAB in the sample and population-wide estimates in the radio repair and mechanics samples showed a striking decline as experience increased. In both samples, the appropriate aptitude composite (Electronics Repair for radio repairers and Mechanical Maintenance for mechanics) had high predictive validity for people with up to 2 years in the Marine Corps (population-wide estimates of .69 for radio repairers and .72 for mechanics with 2 to 14 months of service and .52 for mechanics with 15 to 25 months of service). For people with more than 2 years of service, the validity coefficients dropped dramatically—down to .00 for radio repairers, to .15 for mechanics with 26 to 34 months, and to -.07 for mechanics with 35 to 60 months of service. For the mechanics, the validity coefficients showed a steady decline as experience increased. The pattern of declining validity coefficients also held for the riflemen, but it was not as pronounced.

These results have profound implications for the validation of qualification standards. If the zero or negative validity of ASVAB aptitude composites for predicting hands-on test scores were found for other specialties and were

believed to reflect reality, the conclusions could be that job experience can compensate for aptitude and that aptitude standards can be lowered or abolished altogether. One could, however, also argue that because the ASVAB did predict performance during the first 2 years of service, the aptitude standards are justified. The length of time the services are willing to wait for proficiency to mature in a specialty versus the cost of procuring higher-aptitude recruits would need to be weighed through a cost-benefit analysis.

TABLE 7

VALIDITY OF ASVAB FOR PREDICTING HANDS-ON TEST SCORES, BY LEVEL OF EXPERIENCE

Months in service	Validity coefficients[a]		Standard deviation[b]	Number of cases
	Sample[b]	Population[c]		
Ground Radio Repairers				
15-25	.41**[d]	.69	9.44	38
26-35	.00	.00	8.50	53
36-48	.00	.00	7.50	37
Total	.17*	.37	8.60	128
Automotive Mechanics				
2-14	.54**	.72	12.42	57
15-25	.46**	.52	16.96	56
26-34	.13	.15	17.19	53
35-60	-.05	-.07	13.35	54
Total	.29**	.37	15.16	220
Infantry Riflemen				
3-13	.58**	.73	13.24	42
14-15	.43**	.66	10.83	47
16-26	.35**	.54	11.62	63
Total	.44**	.64	11.87	152

a Validity of appropriate ASVAB aptitude composite
b Sample value
c Population-wide estimated value, based on the univariate model; the population standard deviation equals 20
d Coefficients significant at the 1-percent level are shown by **, those significant at the 5-percent level are shown by *

Another line of argument is that the low predictive validity of the ASVAB for more experienced people is an artifact arising from a low content validity of the hands-on tests. If their content is appropriate for people newly assigned to their first duty station but much less so for people with more experience, then the results make sense and they do not jeopardize qualification standards. This line of argument implies that hands-on tests appropriate for more experienced people can be constructed. Once constructed, an empirical demonstration that the ASVAB does, in fact, predict hands-on performance of experienced people is then required. In the absence of hard empirical data, the argument that the content validity of the hands-on tests is faulty remains speculative.

The implications of these results for the Marine Corps Job Performance Measurement Project are described in section 3.

Other Evaluations of the Content Validity of the Hands-On Tests

Correlations between hands-on test scores and other measures of performance (written proficiency tests, supervisors' ratings of job proficiency, and grades in specialty training courses) are shown in table 8. The correlation coefficients are sample values, which means that they are lower than the population-wide estimates would be, but their pattern is essentially the same. The coefficients that are based on a smaller number than the total in each subgroup are shown with the number of cases in parentheses.

The correlation between the hands-on tests and supervisors' rating tended to increase with experience in all three specialties. Whatever the hands-on tests were measuring for the more experienced people, they had a higher relationship to the proficiency valued by supervisors than they did for the less experienced people. The written proficiency tests and training grades tended to have a consistently positive pattern of relationship to the hands-on tests.

The relationship of hands-on tests to other indicators of performance (reduction in grade, irregular discharge, and promotion rate) is shown in table 9. There is some suggestion that receiving an irregular discharge, which implies less than fully satisfactory behavior, was related to hands-on test scores. People with low scores tended more often to receive irregular discharges. More riflemen with low scores were reduced in grade (busted), but that was not true for mechanics. No radio repairers in the sample were

busted. Promotion rate, which shows the speed with which the examinees were promoted during their first enlistment, was not related to the hands-on test scores in any of the specialties. The relationships in the sample of riflemen are plotted in figure 5.

TABLE 8

CORRELATION OF HANDS-ON TEST SCORES WITH OTHER MEASURES OF PERFORMANCE

Months in service	Correlation coefficients[a] Written test	Supervisors' ratings	Training grades	Number of cases
		Ground Radio Repairers		
15-25	.23	.21(37)	.37(22)[b]	38
26-35	-.05	-.04(47)	-.13(36)	53
36-48	.17	.43(23)*	-	37
Total	.12	.18(108)	.21(63)	135[c]
		Automotive Mechanics		
2-14	.28*	.05	.39(54)**	57
15-25	.34**	.11	.27(52)*	56
26-34	.31*	.44(51)**	.48(30)**	53
35-60	.14	.37(45)**	.24(41)	54
Total	.29**	.24(209)**	.34(177)**	220
		Infantry Riflemen		
3-13	.54**	.17(41)	-	42
14-15	.38**	.31(43)*	-	47
16-26	.46**	.51(58)**	-	63
Total	.45**	.31(142)**	-	152

a. Coefficients significant at the 1-percent level are shown by **; those significant at the 5-percent level are shown by *

b. The number of cases used to compute the correlation, when different from the number in the subgroup, is shown in parentheses

c. Includes seven people with more than 48 months in the Marine Corps

TABLE 9

RELATIONSHIP OF HANDS-ON TEST SCORES TO OTHER INDICATORS OF PERFORMANCE

Hands-on score	Number of cases			Promotion rate[a]
	Total	Reduced in grade	Irregular discharge	
		Ground Radio Repairers		
44-98	43	–	5	.124
100-128	51	–	1	.138
130-140	43	–	1	.133
Total	137	–	7	.127
		Automotive Mechanics		
46-69	57	17	9	.139
70-73	46	8	7	.141
74-76	54	9	4	.137
77-80	63	14	5	.141
Total	220	48	25	.140
		Infantry Riflemen		
48-102	39	20	15	.116
103-138	41	15	9	.139
139-164	41	15	11	.115
165-219	40	7	5	.140
Total	161	57	40	.128

a. Enlisted grade divided by number of months in the Marine Corps.

MEASUREMENT VALIDITY

The measurement validity of the hands-on tests was evaluated by examining differences among scores assigned by the test administrators in the radio repair and automotive mechanics samples and the relationship between test scores and testing dates in all three samples.

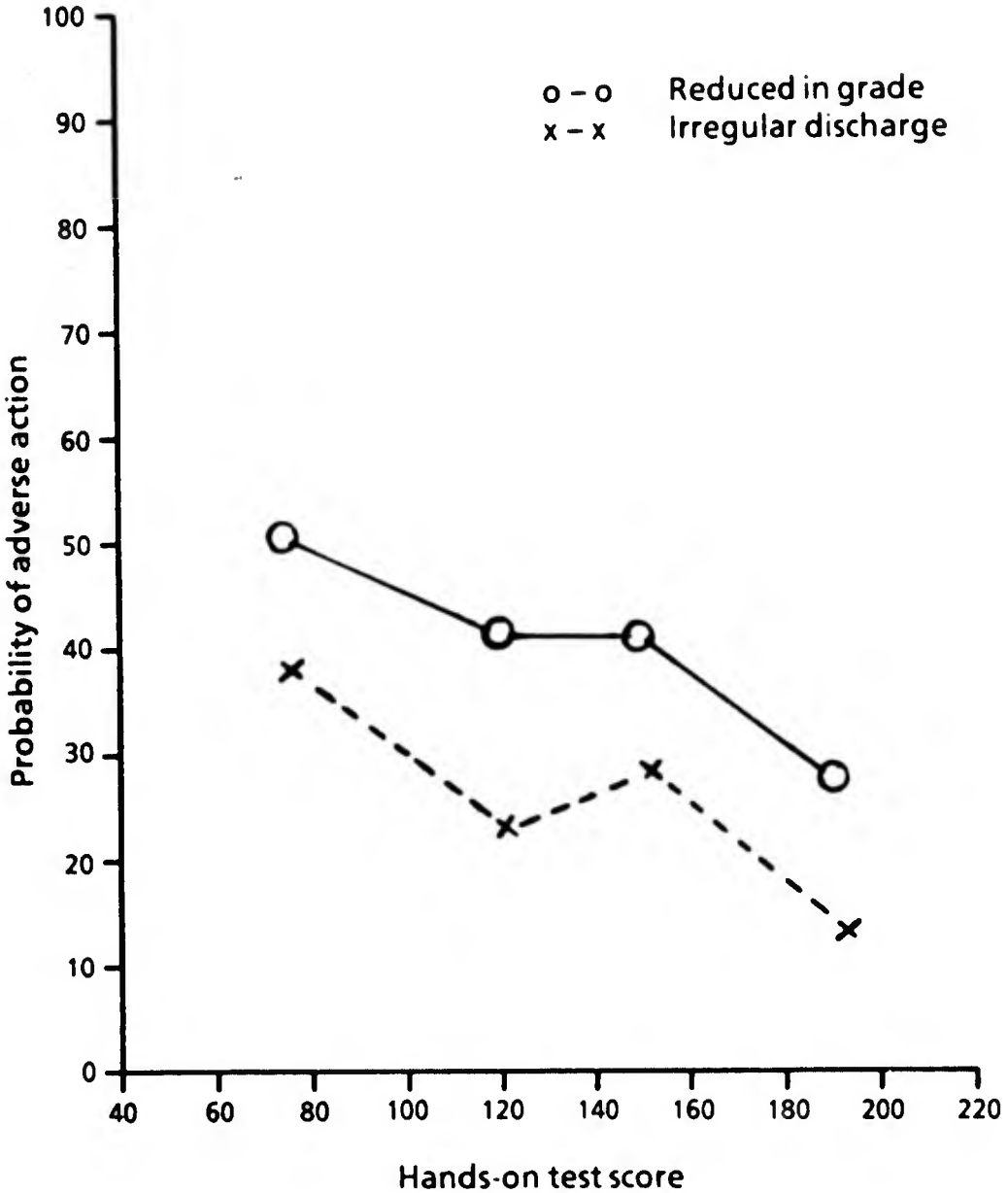


FIG. 5: HANDS-ON TEST SCORES RELATED TO ADVERSE ACTIONS FOR INFANTRY RIFLEMEN

The mean hands-on test scores assigned by individual test administrators are shown in table 10. Most of the differences are statistically significant, and they were not related to differences in the written test or ASVAB [1]. Note that the first 44 Radio Repair examinees did not have their answer sheets signed by the test administrator, and the test scores were lower than those assigned later, when the administrators signed their sheets, except those assigned by administrators 5 and 6.

TABLE 10

MEAN HANDS-ON TEST SCORES ASSIGNED BY TEST ADMINISTRATORS

	1	2	3	4	5	6	Unknown
Radio Repairers							
Mean hands-on score[a]	120	127	115	117	107	70	111
Number of cases	16	13	20	19	20	2	44
	1A	**1B**	**2**	**3**	**4**		
Automotive Mechanics							
Mean hands-on score[b]	75	72	68	74	74		
Number of cases	74	58	64	23	42		

a. Maximum score is 140. Standard error of mean is about 5
b. Maximum score is 80. Standard error of mean is about 1

Administrator 1 for the mechanics tested 132 examinees: 74 before 21 September 1981 (1A) and 58 on and after 21 September (1B). Figure 6 shows the distribution of scores assigned by this administrator in the first and second half of the testing period. During the first half, more than 50 percent of the hands-on test scores assigned were 77 through 80, the top four scores (labeled "very high" in figure 6); during the second half, the percentage of very high scores dropped to less than 20. The percentage of lower scores (including those 46 to 69, labeled "very low"; 70 to 73, labeled "low"; and 74 to 76, labeled "high") was correspondingly greater during the second half of the testing period.

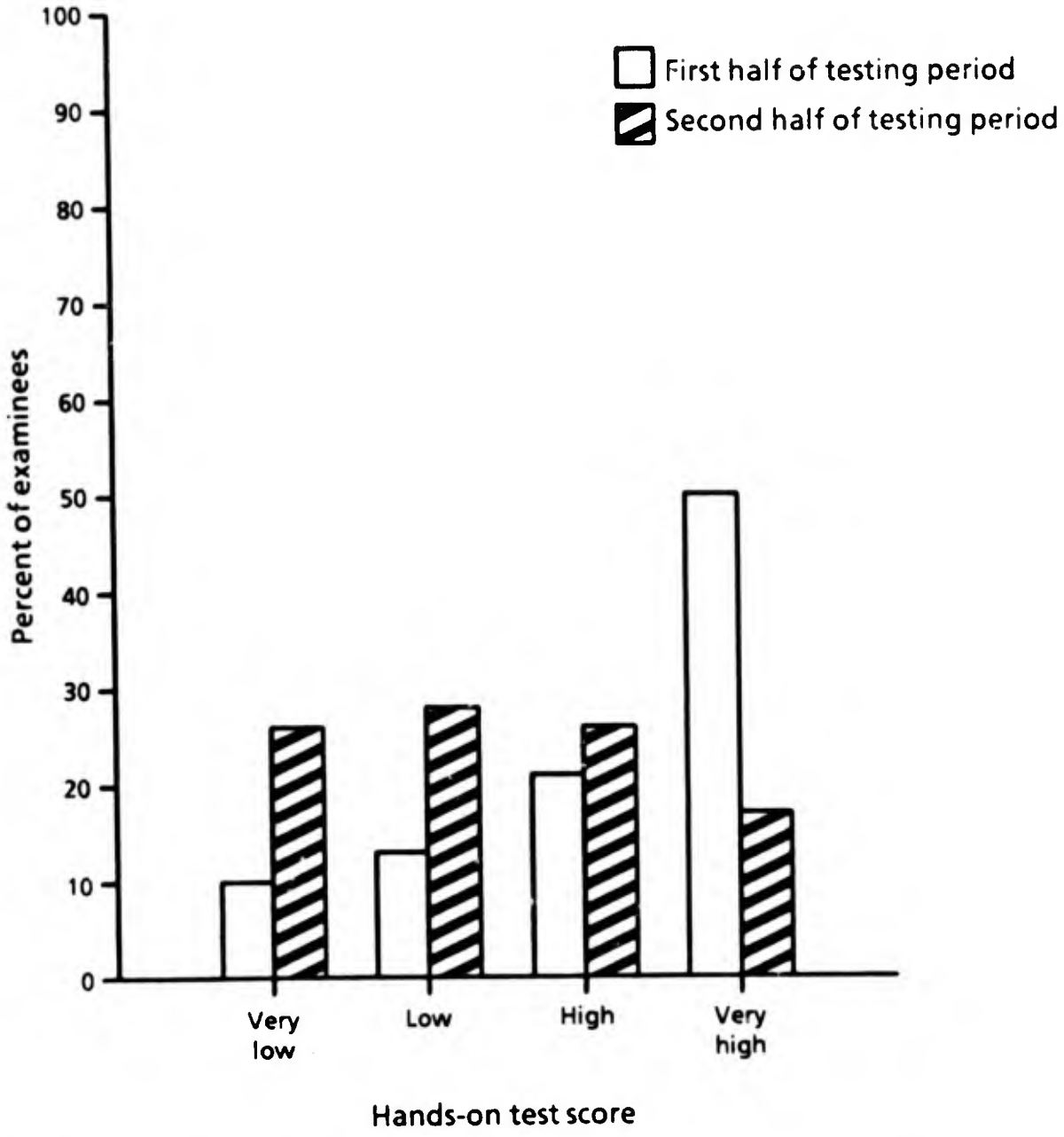


FIG. 6: HANDS-ON TEST SCORES ASSIGNED BY A TEST ADMINISTRATOR TO AUTOMOTIVE MECHANICS

Part of the differences in scoring calibration among the test administrators for radio repairers and mechanics was removed by standardizing the hands-on scores assigned by each administrator to have the same mean and standard deviation. When the assigned scores were standardized, their predictability from ASVAB scores was increased. The validity coefficients of the appropriate ASVAB aptitude composites with original and standardized test scores are as follows:

	Validity coefficient	
Specialty	Original	Standardized
Radio Repair	.28	.41
Auto Mechanic	.34	.44

The validity coefficients were corrected for range restriction using the univariate model.[1]

The population-wide estimates of ASVAB validity for predicting the hands-on test scores as originally assigned by the test administrators were marginal, but the coefficients based on standardized scores are quite respectable. This simple adjustment to standardize the test scores significantly improved the measurement validity of the scores as evidenced by the impact on the size of the ASVAB validity coefficients.

The second component of measurement validity evaluated in this analysis is the effect of testing date on the hands-on test scores. The mean hands-on test scores by testing date are shown in table 11. The number of minutes each examinee took to complete the test (testing time) and the correlation between test scores and testing time are also shown for the radio repairers and mechanics. No testing-time data are shown for the sample of riflemen because testing time was not a meaningful variable for their test. The relationship between test scores, as the percent correct, and testing date is plotted in figure 7.

1. The standard deviation of the Mechanical Maintenance aptitude composite in the mechanics sample was 16.70; the standard deviation of the Electronics Repair composite in the radio repairers sample was 9.04. Both have a standard deviation of 20 in the population.

TABLE 11

RELATIONSHIP BETWEEN HANDS-ON TEST SCORES AND DATE OF TESTING

	Mean		Standard deviation			
Test date	Test score	Testing time (min.)	Test score	Testing time (min.)	Score-testing time correlation	Number of cases
			Ground Radio Repairers			
Aug 9 – Sep 18	101.2	197.1	26.22	24.09	– .43	57
Sep 21 – Oct 9	112.0	181.9	30.99	30.42	– .62	39
Oct 13 – Oct 30	122.1	163.6	19.86	41.86	– .52	41
			Automotive Mechanics			
Aug 3 – Aug 19	73.8	136.3	6.10	23.23	– .28	53
Aug 20 – Sep 18	75.2	141.0	5.20	27.89	– .03	59
Sep 21 – Oct 9	70.5	138.8	7.28	26.99	– .05	59
Oct 13 – Oct 31	69.6	130.6	5.26	20.61	+ .14	49
			Infantry Riflemen			
Jul 15 – Aug 31	144.2	–	31.91	–	–	48
Sep 1 – Sep 30	129.1	–	39.07	–	–	39
Oct 1 – Oct 31	137.6	–	41.05	–	–	36
Nov 1 – Nov 10	122.6	–	44.53	–	–	36

In the radio repair sample, the mean hands-on test score increased and mean testing time decreased in the later testing periods. The results in the mechanics sample, however, were the opposite: test scores decreased after 18 September, and testing time remained essentially unchanged. The hands-on test scores in the riflemen sample showed no consistent trend with test date.

The test administrators for the mechanics were instructed on 21 September to refrain from giving help to examinees, which may explain the change in the mechanics' scores. There is no similar explanation for the opposite shift in radio repairers' scores. These results are consistent with the

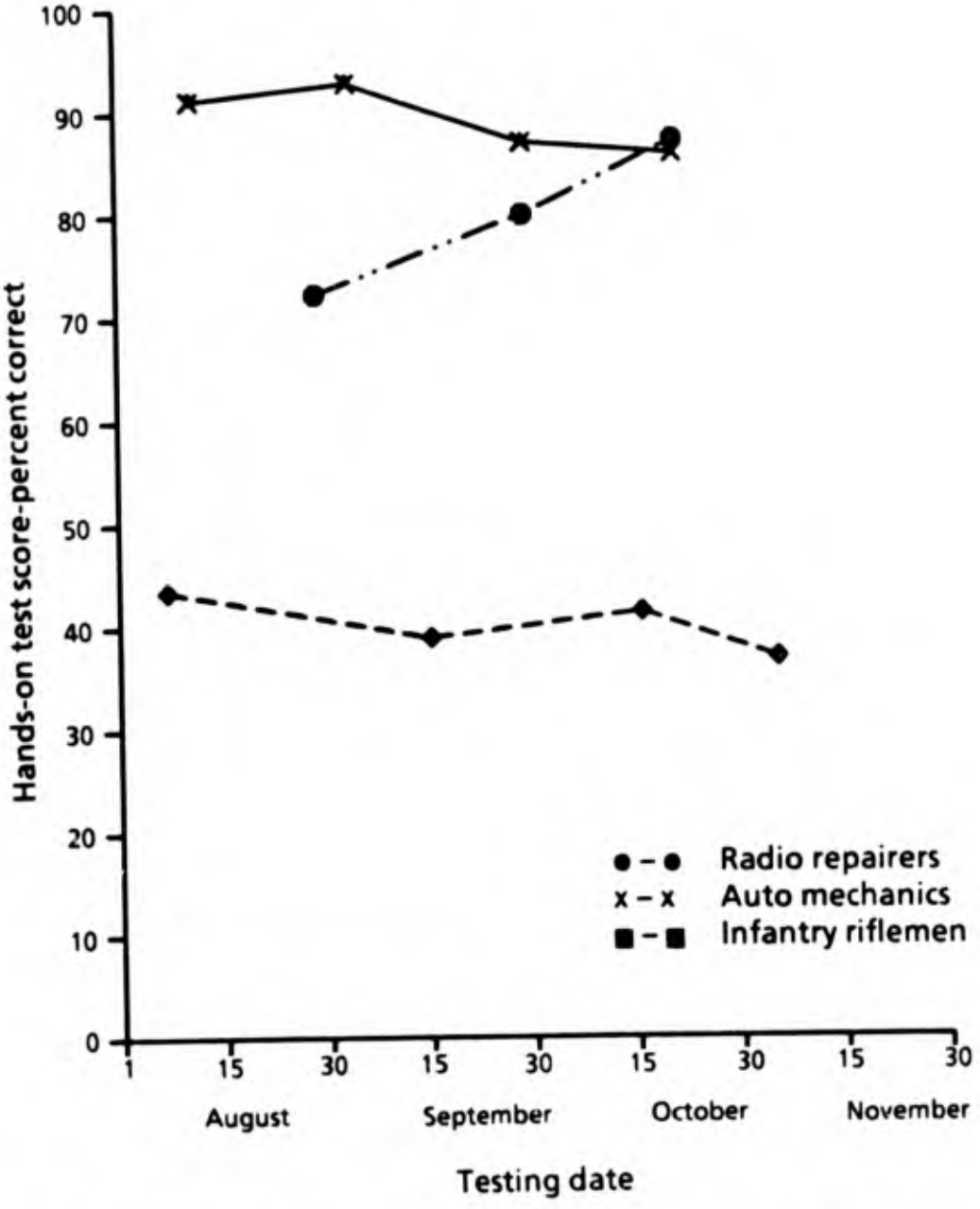


FIG. 7: HANDS-ON TEST SCORES RELATED TO TESTING DATE IN THREE MARINE CORPS SPECIALTIES

repairers' scores, a more plausible explanation is that the scoring calibration of the test administrators shifted. In either event, the hands-on test scores in the samples of radio repairers and mechanics changed systematically over time, and this fact lowers the measurement validity of the tests.

SECTION 3

IMPLICATIONS

SUMMARY OF RESULTS

The purpose of this analysis was to examine the content and measurement validity of hands-on tests developed for three Marine Corps specialties – Ground Radio Repair, Automotive Mechanic, and Infantry Rifleman. Based on the widespread practice among the military services and civilian employers of basing salary and promotion in part on job experience, test scores were expected to increase with time served in the Marine Corps. In this regard, the results for the hands-on test were negative; the scores were essentially unrelated to months of service. For the written proficiency tests in the two technical specialties (Radio Repair and Auto Mechanic), the results were positive; written test scores did increase with time in the Marine Corps. Riflemen in the sample had served 2 years or less, and their test results were inconclusive. The lack of relationship of the hands-on tests to job experience raises questions about test content validity.

As a further check on the content validity of the hands-on tests, the effects of aptitude and experience on job performance as measured by the tests were evaluated. Based on experience with the ASVAB, the expectation is that job performance should increase with both aptitude and experience, and therefore the interaction between the two variables should be negligible. However, for the two technical specialties, the interaction of aptitude and experience had a significant effect on test performance scores. People with high aptitude started with high hands-on test scores and did not improve with experience. In fact, their scores tended to get slightly worse. People with low aptitude, by contrast, started lower, but with experience improved their scores to a level equal to or better than those for people with high aptitude. On the written test, aptitude and experience showed no interaction; scores were uniformly higher for people with high aptitude and longer experience.

The hands-on test scores were found to be related positively to supervisors' ratings of proficiency, especially for the more experienced examinees. There was also a slight correlation between low test scores and reduction in grade and irregular discharge.

The measurement validity of the hands-on test in the two technical specialties was evaluated by comparing the scoring calibration of the test administrators and examining shifts in test scores across testing dates. Test scores were found to differ among administrators and over time; that is, different test administrators used different scoring standards, and the same administrators used different standards at different times.

The results of this analysis raise questions about both the content and the measurement validity of the hands-on tests used in the earlier feasibility study [1]. In the remainder of this section, the implications of these results for validating qualification standards, designing the Marine Corps Job Performance Measurement Project, and institutionalizing job performance measurement are discussed.

VALIDATING QUALIFICATION STANDARDS

The most salient finding from this analysis is the confirmation that the ASVAB is a valid predictor of job performance. Qualification standards therefore are justified. The high predictive validity, however, depends on the level of job experience. But the fact that the validity coefficients are high satisfies professional and legal requirements.

The model relating ASVAB to job performance as measured by hands-on tests may need to be more complex than the simple linear model used in the feasibility study. In that study, performance was assumed to be a linear function of aptitude, and the hands-on test scores for each administrator in the two technical specialties were linearly standardized to have the same mean and standard deviation. The model followed conventional practice in validation studies.

This analysis indicates that complex nonlinear relationships are involved. The relationship of aptitude to hands-on test scores in this analysis was shown to be affected by job experience and by testing date. The form of the nonlinear relationship remains to be worked out, and these data do not provide an adequate basis.

Both content and measurement validity of the hands-on tests need to be, and probably can be, improved. Through a more systematic incorporation of job requirements into the tests and more rigorous quality control over the test administrators, the quality of the performance data can be upgraded. The

formulation of the proper model to account for job performance should be based on data in which the research community and personnel managers have confidence. The efforts of the JPMWG are expected to provide the quality of data required.

DESIGNING THE MARINE CORPS JOB PERFORMANCE MEASUREMENT PROJECT

The initial design for the Marine Corps Job Performance Project called for testing first-term personnel in about eight specialties. After evaluating the results of this analysis, the Marine Corps has developed a new research design.

Instead of testing only one sample in each specialty, the following three samples will be tested:

- New graduates from specialty training courses; up to 600 examinees per specialty
- First-term Marines with the full range of job experience; up to 600 examinees per specialty
- Second-term Marines who are functioning as first-level supervisors; up to 600 examinees per specialty.

The hands-on tests will be carefully developed to cover the full range of job requirements and to permit inferences about changes in proficiency as experience increases. Most of the test content is expected to be common to all three samples. The test administrators will be trained to ensure that their scoring is properly calibrated and continually monitored to ensure that it remains so.

INSTITUTIONALIZING JOB PERFORMANCE MEASUREMENT

The Job Performance Measurement Project was originated because of widespread dissatisfaction with the extant justification for enlistment standards. Historically, the ASVAB has been validated against grades in specialty training courses. Except for isolated efforts, which rarely were designed specifically to validate aptitude tests, military aptitude tests have been validated against surrogate measures of performance. The focus of the

joint-service project is to develop hands-on tests and use them for validating the ASVAB.

In the spring of 1985, the Office of the Secretary of Defense (OSD) initiated efforts to institutionalize job performance measurement. Although hands-on tests are not specified as the measure of job performance, the OSD initiative arose in the context of this project, and the use of hands-on tests is a distinct possibility.

The problems encountered in this analysis with the content and measurement validity of the hands-on tests raise concerns about institutionalizing these tests. Feasibility is a more important consideration in a large scale testing program than in a research project. Unless care is taken, the test content would almost certainly be limited to job requirements that are relatively easy to test in the hands-on mode.

Attaining and maintaining measurement validity of an operational testing program, in which the welfare of individuals is at stake, would be harder than attaining content validity. If examinees knew that their careers could be affected, they would be motivated to learn what is being tested and to practice the test. If test administrators knew that the scores influenced decisions about the examinees or their commanders, they would tend to be more lenient. Hands-on test scores are highly susceptible to manipulation and, like ratings, can be readily turned to obtain a desired result.

Even though the data collected in the Job Performance Measurement Project are intended only for research, the test administrators were found to be lenient. No personnel decisions, such as on promotion or retention, and no evaluations of training programs or units were based on the data. No one had a vested interest in the scores assigned to individuals or to groups. If test administrators without any incentives to make examinees look good were lenient, how much more would real consequences for people cause the hands-on scores to be inflated?

A reasonable alternative to hands-on testing is to institutionalize a surrogate measure of performance that has been shown to have content validity, is economical to develop and administer, is relatively immune to scoring manipulation by test administrators, and can be kept secure. At this stage, no one has enough data to know the form of a valid surrogate. Assuming that the Job Performance Measurement Project will not be

replicated, perhaps now is the time to evaluate whether the combined efforts of the services are likely to result in measures worthy of institutionalizing.

CONCLUSIONS

- The ASVAB is a valid predictor of job performance as measured by hands-on tests.
- But hands-on tests lack robustness; that is,
  - Content validity is sensitive to job experience.
  - Measurement validity is sensitive to the scoring calibration of test administrators.
- Complex models are required to establish the relationship between aptitude and hands-on tests.

REFERENCES

[1] CNA Report 89, *An Evaluation of Using Job Performance Tests To Validate ASVAB Qualification Standards*, by Milton H. Maier and Catherine M. Hiatt, May 1984

[2] Navy Personnel Research and Development Center, NPRDC TN 82-20, *Marine Corps Job Performance Test for Three Enlisted Specialties*, by David J. Chesler, Chester R. Belinski, and Marc A. Hamovitch, Jun 1982

APPENDIX A

MEAN PERFORMANCE TEST SCORES BY LEVEL OF EXPERIENCE

APPENDIX A

MEAN PERFORMANCE TEST SCORES BY LEVEL OF EXPERIENCE

The mean performance test scores for each specialty at each level of job experience are shown in table A-1. In table A-2, the samples have been divided into low and high levels of aptitude.

TABLE A-1

MEAN PERFORMANCE MEASURE SCORES[a] BY LEVEL OF EXPERIENCE

		Performance measure					
		Hands-on test		Efficiency score		Written test	
Months in service	Number of cases	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
			Ground Radio Repairers				
15-25	38	48.5	12.26	48.0	8.45	46.1	9 28
26-35	53	50.0	8.33	50.6	10.56	50.5	10.40
36-48	37	52.9	9.82	51.9	11.16	53.4	9 25
49-60	9	48.6	8.96	46.7	5.66	56.6	9 42
Total	137	50.2	10.03	50.0	10.00	50.4	10.14
			Automotive Mechanics				
2-14	57	47.6	11.62	48.2	10.05	47.2	9.18
15-25	56	50.0	8.92	50.2	10.57	49.9	9 81
26-34	53	50.3	10.24	50.0	10.02	49.5	10.11
35-60	54	53.0	8.09	51.4	9.27	54.5	8.26
Total	220	50.2	9.95	50.0	10.00	50.2	9 68
			Infantry Riflemen				
3-13	42	51.2	9.89	–	–	48.0	10.32
14-15	47	48.3	9.28	–	–	49.2	8.29
16-26	63	49.6	10.12	–	–	50.7	8.20
Total	152	49.6	9 80	–	–	49.5	8.87

a Standardized to have a mean of 50 and a standard deviation of 10

TABLE A-2

MEAN PERFORMANCE MEASURE SCORES[a] BY LEVEL OF EXPERIENCE AND APTITUDE

			Performance measure					
			Hands-on test		Efficiency score		Written test	
Apitude level	Months in service	Number of cases	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
			Ground Radio Repairers					
Low	15-25	24	45.5	12.97	45.9	7.97	43.4	9.30
Low	26-35	27	50.2	7.39	49.2	7.18	46.6	9.31
Low	36-48	13	55.3	8.12	54.8	15.00	46.2	8.05
High	15-25	14	53.6	9.25	51.6	8.31	50.6	7.55
High	26-35	26	49.8	9.36	52.2	13.17	54.5	10.09
High	36-48	24	51.6	10.55	50.0	8.39	57.4	7.37
			Automotive Mechanics					
Low	2-14	34	44.1	12.72	44.7	10.41	45.0	9.80
Low	15-25	27	47.2	11.14	45.2	9.63	45.3	9.30
Low	26-34	24	49.0	10.33	49.0	9.91	46.8	11.93
Low	35-60	26	53.4	6.26	52.8	9.91	51.9	9.42
High	2-14	23	52.5	7.60	53.7	7.17	50.4	7.18
High	15-25	29	52.7	5.08	55.2	9.49	54.3	8.25
High	26-34	29	51.3	10.23	51.8	10.38	51.6	7.93
High	35-60	28	52.6	9.59	50.6	8.96	57.0	6.22
			Infantry Riflemen					
Low	3-13	22	46.9	8.05	–	–	43.3	9.31
Low	14-15	22	45.3	8.20	–	–	44.5	7.23
Low	16-26	32	47.8	10.30	–	–	49.3	7.06
High	3-13	20	55.9	9.72	–	–	53.2	8.93
High	14-15	25	51.0	9.53	–	–	53.3	6.91
High	16-26	31	51.4	9.78	–	–	52.1	9.12

a. Standardized to have a mean of 50 and a standard deviation of 10.