ARI TECHNICAL REPORT

Evaluation of Prototype Job Performance Tests for the U.S. Army Infantryman

by

Michael R. McCluskey, Jules C. Trepagnier, Jr.,
Fred K. Cleary, and James M. Tripp

HUMAN RESOURCES RESEARCH ORGANIZATION
300 North Washington Street
Alexandria, Virginia 22314

OCTOBER 1975

Final Report-CD-(C)-75-9

Prepared for

U.S. ARMY RESEARCH INSTITUTE
for the BEHAVIORAL and SOCIAL SCIENCES
1300 Wilson Boulevard
Arlington, Virginia 22209

The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: HumRRO-FR-CD(C)-75-9
2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)*: EVALUATION OF PROTOTYPE JOB PERFORMANCE TESTS FOR THE U.S. ARMY INFANTRYMAN
5. TYPE OF REPORT & PERIOD COVERED: Final Report. July 1974 - October 1975
6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(S): Michael R. McCluskey, Jules C. Trepagnier, Jr., Fred K. Cleary, and James M. Tripp
8. CONTRACT OR GRANT NUMBER(S): DAHC 19-74-C-0043
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Human Resources Research Organization (HumRRO), 300 North Washington Street, Alexandria, Virginia 22314
10. PROGRAM ELEMENT PROJECT TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: U.S. Army Research Institute for the Behavioral and Social Sciences, 1300 Wilson Boulevard, Arlington, Virginia 22209
12. REPORT DATE: October 1975
13. NUMBER OF PAGES: 41 (12) 44p.
14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*
15. SECURITY CLASS. *(of this report)*: Unclassified
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

Research was performed by the Columbus Office of the HumRRO Central Division (formerly HumRRO Division No. 4).
Project PERFORM

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Combat arms	Performance objectives
Critical tasks	Performance tests
Field tests	Task inventories
Occupational specialties	Training literature

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

In the first phase of the research, 1,885 performance objectives were developed for eight key combat arms MOSs. Phase II of the project was concerned with a field investigation of four prototype performance tests. The results indicated considerable variability between different Test Administrators in terms of their test site preparation and administrative procedures. A low degree of interrater reliability was found for approximately 37 percent of the performance measures in each test. In general, the perceived face validity of each test was high in terms of job relevance and fairness.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified 405 260

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

SUMMARY

INTRODUCTION

This report presents the results of a project conducted in support of work concerned with development of the U.S. Army's Enlisted Personnel Management System (EPMS). The project was conducted in two independent but somewhat related phases. Phase I was devoted to development of performance objectives for critical or important tasks of eight combat arms MOSs. Phase II was devoted to a field investigation designed to determine the potential reliability, validity, and feasibility of prototype performance tests developed from the performance objectives.

PHASE II

The overall purpose of Phase II was to conduct a field evaluation of four prototype performance tests. The specific objectives were as follows:

1. Identify differences in performance test situations established by different Test Administrators in terms of test site preparation and test administration.
2. Determine the degree of interrater reliability associated with observing and scoring performance.
3. Determine the face validity of performance tests in terms of perceived job relevance and fairness.

A total of 100 subjects, consisting of Test Administrators, Observers, and Examinees, were utilized in the field test. The Test Administrators were responsible for the site preparation and administration of all tests and the Observers were required to independently rate the performance of each Examinee on each test. The field test was completed over a two-week period, with each of five groups participating on two consecutive days. Each of the five groups consisted of one Test Administrator, four Observers, and 15 Examinees. During the two-day period, each Examinee completed four performance tests, and was rated by the four Observers and the Test Administrator.

All performance tests were conducted in a field environment. The prototype performance tests were developed in the following areas: LAW, Camouflage, Lifesaving, and Claymore mine. Each test was administered to one Examinee at a time, and the administration times generally ranged from 10 to 20 minutes.

The results indicated there was considerable variability between Test Administrators in terms of test site preparation and administrative procedures. The variability in test sites was generally concerned with providing additional items of equipment or failure to follow the site preparation instructions. Differences in administrative procedures were found in the areas of coaching and providing feedback to the Examinees, verbatim reading of instructions to the Examinee, and various deviations from the test scenario. In terms of the degree of interrater reliability between independent raters, approximately 35 percent of the performance measures for each test were found to be unreliable because of the low agreement between raters. In general, the perceived face validity of each performance test was found to be high in terms of job relevance and fairness or equity. A substantial discrepancy was found, however, between the actual test results and the Examinees' perceptions of whether or not they passed the test.

In order to most effectively implement hands-on performance testing in the MOS evaluations under the Army's Enlisted Personnel Management System (EPMS), the following actions appear indicated:

1. Improve and refine the process of test development, with special emphasis upon specification of performance measures and detailed descriptions of all required procedures for test administration.
2. Provide detailed instructions and extensive training for Test Administrators.
3. Conduct a field test for each candidate performance test, revise it, and retest it prior to implementation.

PREFACE

The research described in this report was conducted by the Human Resources Research Organization (HumRRO) under Contract DAHC19-74-C-0043 with the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI). The research was a joint effort between the U.S. Army Infantry School, U.S. Army Armor School, U.S. Army Field Artillery School, U.S. Army Air Defense School, U.S. Army Combat Arms Training Board (USACATB), U.S. Army Research Institute/Human Research Unit at Fort Benning, Georgia, and HumRRO.

Dr. Milton H. Maier (ARI, Arlington, Virginia) served as the Contracting Officer's Technical Representative for the entire project. Designated points of contact and coordination within ARI were Dr. James A. Caviness (Fort Benning) and Dr. Douglas L. Young (Arlington, Virginia). The Columbus Office of the HumRRO Central Division (formerly HumRRO Division No. 4) was responsible for overall project planning and execution. Dr. T.O. Jacobs was the Director when the research was begun. Dr. Wallace W. Prophet is the current Director of the Central Division and Dr. Joseph A. Olmstead is the Columbus Office Director. Mr. Michael R. McCluskey served as Principal Investigator of the project. The U.S. Army Combat Arms Training Board (USACATB) was the military sponsor of the research and COL Franklin A. Hart was President of the Board. The original Project Officer at USACATB was MAJ Edgar D. Maddox, who was solely responsible for initiating the research requirement which led to the completion of this project. Succeeding Project Officers at USACATB were LTC William Valen and CPT Walter Nakano.

Three HumRRO research offices participated in the Phase I activities. In addition to overall planning and management, the Columbus Office of the Central Division developed performance objectives for the Infantry School and the Field Artillery School. Mr. Jules C. Trepagnier, Jr. developed the vast majority of these objectives for both schools. He was assisted, in part, by Mr. George J. Magner, Mr. James M. Tripp, Mr. Jeffery L. Maxey, and Mr. Fred K. Cleary. The performance objectives for the Armor School were developed principally by Mr. James H. Harris, with additional input

from Mr. Roy C. Campbell and Mr. J. Patrick Ford. These activities of the HumRRO Central Division-Louisville Office were completed under the general supervision of Mr. William C. Osborn, Office Director. The performance objectives for the Air Defense School were developed primarily by Mr. Paul Hermann under the direct supervision of Dr. E.W. Frederickson at the HumRRO Western Division-El Paso Office of which Dr. R.D. Baldwin is Office Director.

The research activities of Phase II were completed by the HumRRO Central Division-Columbus Office and the ARI Human Research Unit at Fort Benning. Mr. Jules C. Trepagnier, Jr. and Mr. James M. Tripp (HumRRO) were responsible for the overall coordination and execution of the field investigation. LTC Robert G. Matheson, Chief of the ARI Human Research Unit at Fort Benning, provided the following personnel for assistance during the preparation and data collection phases of the field test: SSG John E. Lang, SP4 Keith L. Evans, and SP4 William W. Fox.

Meredith P. Crawford
President
Human Resources Research Organization

TABLE OF CONTENTS

Tables

Page

Tables (continued)

EVALUATION OF PROTOTYPE JOB PERFORMANCE TESTS
FOR THE U.S. ARMY INFANTRYMAN

INTRODUCTION

This report presents the results of a project conducted in support of work concerned with development of the U.S. Army's Enlisted Personnel Management System (EPMS). The project was conducted in two independent but somewhat related phases. Phase I was devoted to development of performance objectives for critical or important tasks of eight combat arms MOSs. Phase II was devoted to a field investigation designed to determine the potential reliability, validity, and feasibility of prototype performance tests and to identify variables in the development and administration of performance tests that will impact upon their validity and reliability.

OBJECTIVES OF PHASE II

Several significant changes have been made recently in the Army's Enlisted Personnel Management System. One of these modifications is the inclusion of hands-on performance testing in the annual MOS tests. There are numerous research questions that should be answered before hands-on performance tests are used for this purpose.

The basic purpose of Phase II of the research was to conduct a field investigation of prototype performance tests. The field test was designed to provide initial information concerning the following questions and problem areas:

- The first test objective dealt with test standardization in terms of test site construction and test administration. Five NCOs, with comparable backgrounds and experience, were given the same instructions and materials; the objective was to identify differences in the test sites that were established and differences in the administrative procedures.
- The second objective was concerned with the degree of agreement between independent raters observing the same performance. During the field test, five observers rated the performance of each examinee, and the level of interrater reliability was measured.

- The third objective was to determine the face validity of each prototype performance test. In this study, face validity was defined as the extent to which the performance tests were perceived to be job relevant and considered to be fair and reasonable. All participants in the field test responded to a questionnaire addressing these issues.

ENLISTED PERSONNEL MANAGEMENT SYSTEM

A significant aspect of the Enlisted Personnel Management System is concerned with establishing direct relationships between training, evaluation, and career progression. Five levels of training and the corresponding skill levels have been established and approved. This partitioning of training and skill level classifications will permit finer discriminations in professional development training and evaluation. The training and evaluation of enlisted personnel will be more meaningfully tied to career progression through the new Skill Qualification Test (SQT). The scores on the SQT will be used to either verify current skill level or indicate attainment of the next higher skill level. These decisions concerning skill levels will have a direct impact on enlisted personnel classification, evaluation, promotion, and training.

The present concept of the SQT specifies the following test components: (1) written test items, (2) hands-on performance test items, and (3) task areas where proficiency levels will be certified by the commander. Since the SQT will be used as the basic input for making crucial personnel decisions, it must be determined that all components of the performance objectives are valid with respect to actual job requirements. In addition, the performance test items derived from these objectives must be reliable, valid, and provide standardized evaluation across the Army. The basic purpose of Phase II of the current project was to provide initial information concerning some of the questions and anticipated problems in the areas of scorer reliability, standardization in terms of test site construction and test administration, and face validity.

METHOD

SUBJECTS

Subjects used during the test were members of the 197th Infantry Brigade, Fort Benning, Georgia. The brigade customarily assigns support functions to a battalion-size organization on a rotational basis. During the period of the field test, it was the responsibility of the 3d Battalion, 7th Infantry, to complete support functions.

The test plan specified that 100 subjects in the grades of E5-E7 with MOS 11B and a 40-skill level would be required. Of this number, 75 subjects (E5-E6) would serve as Examinees, 20 (E6-E7) would serve as Observers, and five (E6-E7) would serve as Test Administrators.

It was believed that Examinees meeting these criteria would be well qualified to complete the performance tests, and would be capable of providing cogent input concerning face validity of the tests. A sufficient number of Examinees with these qualifications were not available, however, and it was necessary to include Examinees with a 20-skill level and MOS 11B or 11C.

It was requested that Test Administrators and Observers be subjects who were familiar with the procedures and problems associated with training and testing at the unit level but not particularly experienced in performance testing. The Test Administrator was to be senior in grade or duty position to his assigned Observers, while both these categories were to have rank equal to or greater than Examinees with whom they were working.

The characteristics of the actual samples obtained for the Test Administrators, Observers, and Examinees are shown in Table 1. Four of the Test Administrators were in grade E7, and the other held the grade of E6. All Test Administrators had a skill level of 40; three had a Light Weapons Infantryman Primary MOS, one had an Armor Reconnaissance Specialist Primary MOS, and one had an Airplane Repairman Primary MOS. Each Test Administrator was senior to his assigned Observers, either in grade or time in grade. Each Test Administrator and Observer was senior in grade or equal in grade to the Examinees they were testing or rating.

Table 1

TEST ADMINISTRATOR, OBSERVER, AND EXAMINEE GRADE AND PRIMARY MOS DATA[1]

	Grade						Primary MOS								
	E2	E3	E4	E5	E6	E7	11B40	11B20	11B10	11C20	11C40	11D40	11H20	17E40	67G40
GROUP A															
Test Administrator						1	1								
Observer				1	3		3	1							
Examinee			12	3			3	9	2	1					
GROUP B															
Test Administrator						1					1				
Observer					4		3				1				
Examinee			12	3			2	7	3	2	1				
GROUP C															
Test Administrator						1	1								
Observer				2	2		3					1			
Examinee			8	7			6	7			1			1	
GROUP D															
Test Administrator						1	1								
Observer				1	3		3				1				
Examinee	1	1	13					12	1	1			1		
GROUP E															
Test Administrator					1										1
Observer				1	3		4								
Examinee		2	9	3			3	6	3	1			1		

MOS Legend:

11B - Light Weapons Infantryman
11C - Infantry Indirect Fire Crewman
11D - Armor Reconnaissance Specialist
11H - Infantry Direct Fire Crewman
17E - Field Illumination Crewman
67G - Airplane Repairman

[1]For Examinees, data reflects that information gathered on Day 1 of each test cycle; it does not reflect changes as a result of replacing one Examinee with another.

Seventy-five percent of the Observers were in grade E6, with the remaining 25 percent in grade E5. Ninety-five percent of the Observers had a skill level of 40 and the others had a skill level of 20. All Observers had an Infantry-type Primary MOS.

The Examinees ranged from grade E2 through grade E5, with a multitude of MOS and skill levels represented. The most common grade was that of E4, which made up 72.9 percent of the group; the next largest group by grade was that of E5, which made up 21.6 percent of the group. Ninety-nine percent of the Examinees had an Infantry-type Primary MOS, and 88 percent had a skill level of 20 or greater.

During the field test, 8.1 percent of Examinee participants were replaced by other men, and 18 Examinees were classified as "No Shows." A "No Show" was defined as an Examinee who failed to take one performance test.

The members (Test Administrators, Observers, and Examinees) of each group for Groups A-D were selected from the same company; subjects in Groups A and B were from Company B, those in Group C were selected from the Combat Support Company, and those in Group D were drawn from Company A. Group E, on the other hand, was composed of a composite group representing Headquarters and Headquarters Company, Companies A and B, and Combat Support Company. Therefore, all types of personnel in Groups A-D were quite familiar with each other, and the Test Administrators and Observers judged the performance of their own subordinates during the test.

TEST LOCATION

The field test was conducted at Fort Benning, Georgia, in a 197th Infantry Brigade training area which was about one kilometer square. The terrain in the test area satisfied the conditions prescribed by the performance tests, and was located close to the barracks and headquarters of subjects as a secondary consideration.

PERFORMANCE TESTS EVALUATED

The overall purpose of the field test was to surface as many problems and issues as possible concerning the development and administration of hands-on performance tests. In order to identify potential problem areas that the TRADOC schools might have in test development, the performance tests used in this investigation were furnished by the Infantry School. The Office of Directorate of Doctrine and Training Development (DDTD), USAIS, developed four prototype performance tests for use during the field test. This action was undertaken by DDTD even though their planned production of such items was approximately ten months distant. Because of the short time period available for test development, the resulting performance tests did not represent the official position of the Infantry School with respect to test content or format.

Four job tasks were selected by HumRRO for the development of performance tests by DDTD. The performance tests covered the following subject areas:

Test Item No. 1: Prepare M72A2 LAW for Firing; Restore M72A2 LAW to Carrying Configuration

Test Item No. 2: Install/Recover an Electrically Armed Claymore Mine

Test Item No. 3: Apply the Four Life-Saving Measures (clear the airway, stop the bleeding, treat for shock, protect the wound)

Test Item No. 4: Camouflage Yourself and Your Individual Weapon

The tasks selected for performance test development were expected to provide a wide variety of testing situations and problems. Some of the characteristics of the tasks and resulting tests are given below.

1. The tasks were common to the eight MOSs represented in Project PERFORM.
2. Soldiers with a 10 skill level should be able to perform them.
3. They made no unrealistic demands for terrain, equipment, or materials.

4. It appeared that the instruments would offer great latitude in the selection of precise test sites, and in the organization of these sites for testing.

5. All tests required a "hands on" performance by Examinees.

6. Tests 1 and 3 required continuous, close, and precise observation by the Test Administrators and Observers during the entire performance by Examinees.

7. Tests 2 and 4 were evaluations of the end product of Examinee performance.

8. Test 3 utilized simulation in the test situation.

9. It appeared that the performance measures would require various degrees of judgment by the Test Administrators and Observers.

The performance test items used by HumRRO during the field test are shown in Appendix A.* The test items are the products of the service school with slight modifications made by HumRRO to fit them more closely to the needs of the field test.

Some elements of a task were omitted when the total time required would exceed that available for testing. A maximum of 15 minutes could be allocated to each task; ten minutes was to be devoted to actual testing and five minutes to administrative procedures before and after each test. For 15 Examinees, this required three hours and 45 minutes for testing when the schedule was strictly followed. Where modification of tests were believed necessary, complete and independent elements were omitted. For example, the camouflage test was changed from one which required camouflage of exposed skin, weapon, load-bearing equipment, uniform, and helmet cover to one which required camouflage of exposed skin and weapon, with a modified time standard which was coordinated with DDTD.

EQUIPMENT

The equipment and materials shown in Table 2 were made available to Test Administrators for conducting the performance tests. During the development of the test site and situation, the Test Administrators

---

* Appendixes are available from ARI.

Table 2

EQUIPMENT AND MATERIALS LISTED IN PERFORMANCE TESTS

Test Item	Equipment/Material	Number
Prepare M72A2 LAW for Firing; Restore M72A2 LAW to Carrying Configuration	Stopwatch	1
	Warsaw Pact Armored Vehicle Silhouette	1
	M72A2 LAW (inert)	2 per test site
	M72A2 LAW [with cracked tube] (inert)	1 per test site
Install/Recover an Electrically Armed Claymore Mine	Stopwatch	1
	Measuring Tape (marked in meters)	1 (50-meter length)
	M18A1 Anitpersonnel Mine (inert), complete with M7 Bandoleer, M57 Firing Device, M40 Test Set, Firing Wire, and Blasting Cap (inert)	2
Apply the Four Life-Saving Measures (clear the airway, stop the bleeding, treat for shock, protect the wound)	Stopwatch	1
	Manikin, Resuscitation Training (FSN 6910-782-558)	I
	Moulage (open leg wound)	1
	Fatigue Trousers	1 pair
	Fatigue Shirt	1
	Web Belt	1
	Load-Bearing Equipment, complete with poncho	1 set
	Helmet, steel (or helmet liner)	1
	Boots	1 pair
	Field First Aid Dressing	1 per Examinee
Camouflage Yourself and Your Individual Weapon	Stopwatch	1
	Water	10 gal. per test site
	Burlap Garnishing Strip	20 ft. per Examinee
	Camouflage Paint Stick (appropriate to background vegetation)	1 per 20 Examinees

frequently requested additional materials or items of equipment. The additional items requested were compiled by task and are shown in Appendix B.* The additional materials and equipment were not used by every Test Administrator, but were used one time or more by at least one Test Administrator.

The Test Administrators were not required to obtain the aforementioned equipment/materials themselves, but rather HumRRO insured their availability through coordination with the U.S. Army Infantry Human Research Unit (USAIHRU), which delivered the items to test sites each day. Each Test Administrator was informed that requests for additional materials required for the conduct of any test should be submitted to the HumRRO or USAIHRU point of contact as soon as possible.

PROCEDURE

Coordination

The field test was conducted during the periods 14-18 July and 21-25 July 1975. During the four weeks preceding the test, coordinations were made with the Office of Directorate of Doctrine and Training Development; USAIS; ARI; the Infantry Human Research Unit; and Office S3, 197th Infantry Brigade. These coordination efforts addressed the preparation of prototype performance tests, experimental design, data collection, statistical analyses, and material, terrain, and personnel support requirements.

Organization of Subjects

Subjects were organized into five groups, designated Group A-E, respectively. Each group was composed of the following subjects by type and number: Test Administrator - 1; Observer - 4; and Examinee - 15.

The Headquarters, 3d Battalion, 7th Infantry, 197th Infantry Brigade, appointed a Project Officer to coordinate and control the selection of subjects, and to insure that they arrived on-site in accordance with the test plan. The Test Administrators and Observers were made available two days prior to the conduct of tests they were scheduled to administer and

---

* Appendixes are available from ARI.

were on-site during the additional two days of testing. The Examinees were made available for two days of testing, and they were scheduled to arrive at the test site every 30 minutes in groups of three men each.

Briefing of Test Administrator/Observers

For each group, the Test Administrator and Observers were briefed by HumRRO research personnel two working days prior to their participation in the field test. Each briefing session followed a prepared narrative, which was published in booklet form and issued to the Test Administrator and the Observers during the briefing (Appendix C).* This procedure provided standardization in the briefing of each group and served as a complete record of events for each subject. The following information was contained in each booklet:

Administrative Instructions. Instructions to HumRRO briefer.

General. Background information on the HumRRO organization and Project PERFORM.

Importance of Field Test to Test Administrator/Observers. The relationship between the objectives of the field test and future promotion, training, and classification of noncommissioned officers.

General Test Procedure. An explanation of the "what" and "how" aspects of the field test.

Duties of the Test Administrator. Specification of responsibilities with respect to test site selection, site preparation, and administration of four performance tests to individual soldiers.

Performance Tests. Complete copies of each performance test. Instructions were given that these documents and their references would be the sole basis for organizing and administering the performance tests. (These tests are presented in Appendix A.)*

Duties of Observers. A description of the responsibilities for observing the performance of each Examinee, independently evaluating the performance of the task, and recording the results of observations on a scoresheet or checklist.

---

* Appendixes are available from ARI.

Schedule of Performance Tests. A schedule by test item for each two-day cycle of testing. The schedule was as follows:

Day One - Items 1 and 2	Day Two - Items 3 and 4
14 July	15 July
16 July	17 July
18 July	21 July
22 July	23 July
24 July	25 July

Testing was scheduled to begin at 0800 hours for morning sessions and at 1315 hours for afternoon sessions.

Test Sites. Within the test area allocated to the project, Test Administrators were required to select precise test sites to be used, using guidance provided in each performance test. The Test Administrators were prohibited from using test sites previously prepared by other Administrators.

Equipment Required. The ADMINISTRATIVE REQUIREMENTS section of each performance test outlined the equipment/materials needed for the test. These items were placed on-site one hour prior to test time. Requests for additional equipment/materials, believed necessary for conduct of the test by the Test Administrator, were to be submitted as soon as possible to designated HumRRO or USAIHRU personnel.

Administration of Questionnaires. Information concerning the administration of questionnaires to Examinees, Observers, and Test Administrators, and the procedure for this data gathering effort. Personnel assigned to the USAIHRU were responsible for administering the questionnaires.

Administrative Test Precautions. Instructions that the Test Administrators should not discuss aspects of the field test with individuals other than HumRRO personnel. The Observers were not to discuss their judgments with other Observers, either before, during, or after tests while on-site.

HumRRO Field Test Point of Contact. The names and telephone numbers of HumRRO points of contact, as well as the name and telephone number of the USAIHRU supply point of contact and the 197th Infantry Brigade Project Officer.

Notification of Examinees

One week prior to the test, Examinees were notified by letter (Appendix D)* that they had been selected to participate in the PERFORM field test. They were informed of the four performance test areas by subject, but were not given an outline of the precise requirements or expected performance measures. In addition, the letter covered the following areas:

1. That each test would be graded by noncommissioned officers from their own organization.
2. That the results of the tests would be handled in a confidential manner.
3. Background of the HumRRO organization.
4. Background of Project PERFORM.
5. Importance of the field test to soldiers in the future.
6. Purpose of the field test.

Test Administration Procedure

Testing was scheduled in cycles of two consecutive working days, during the period 14-18 July and 21-25 July 1975. During the first day of any cycle, soldiers were tested on two tasks. During the second day, these same soldiers were tested on two different tasks.

Initially, Test Items 1 and 2 were scheduled for Day One, and Test Items 3 and 4 scheduled for Day Two. This plan was followed for Groups A and B. However, because soldiers in Groups A and B experienced considerable difficulty with Test Item 2, the sequence of test items was changed for Groups C-E. For these groups, Test Items 1 and 4 were scheduled for Day One, and Test Items 2 and 3 scheduled for Day Two.

In conjunction with the sequence change above, a change was also made in one of the Conditions of Test Item 2. The question of which component of the electric wire (blasting cap or shorting plug/dust cover) should be at the free end had not been addressed in the Conditions of the performance test. For Groups A and B, the electric wire for the Claymore had been

---

* Appendixes are available from ARI.

placed on the reel so that its shorting plug and dust cover were at the free end of the wire. It appeared that this was the source of soldier difficulty during emplacement, since these men had been trained only in emplacement from installation site to fighting position (where the blasting cap should be on the free end). The change in the placement of the wire on the reel considerably reduced confusion for Groups C-E.

On each day of any given cycle of testing, Examinees were given performance tests one at a time. These men took one performance test during the morning session and a different test during the afternoon session. Each test was set up and organized by the Test Administrator, and was scored by the Test Administrator and four Observers. During the following day, the same group of Examinees took the remaining two performance tests, which were set up by the same Test Administrator and scored by that Administrator and the same Observers that were used the previous day. This cycle was repeated every two days until all Examinees had been tested.

Equipment and materials needed each day were delivered to the test site one hour prior to test time. The Test Administrator had ample time to organize and arrange the items in accordance with his concept of testing. When the Test Administrator had completed preparation of the site, HumRRO personnel checked the preparations on a procedural checklist specifically designed for the particular test scheduled.

As Examinees arrived on location, they were placed in a ready area away from the test site and assigned a number corresponding to the order of succession. Each Examinee was then cycled to the test site and tested.

Procedural Checklists

A procedural checklist was developed in order to obtain information concerning the extent of variability in performance test administration among the five Test Administrators. A separate checklist was prepared for each prototype performance test (Appendix E)*, with checklist items designed to provide information both unique to each test and common to all four tests. Each checklist was divided into two sections.

---

* Appendixes are available from ARI.

Part I was concerned with pretest preparation on the day of the test. Checklist items were designed to obtain information concerning the degree of similarity between the actual test site conditions and those conditions specified in the performance tests. Specifically, this section dealt with test site characteristics, observer briefings, utilization of equipment, and preparation of scoresheets.

Part II was concerned with the administrative procedures used during the test. Checklist items were designed to provide information concerning the inter- and intravariability of Test Administrator behaviors and actions. This section dealt with the Test Administrator's performance of activities indicated in the performance test, his reading of instructions to Examinees, his coaching of Examinees, and the degree of rapport established with the Examinees.

Performance Test Scoresheets

As each Examinee performed each task, his actions were independently evaluated by the Test Administrator and the four Observers assigned to each group. A Test Administrator/Observer scoresheet was developed for evaluating each Examinee by extracting the Scenario section and the performance measures contained in each of the four performance tests and compiling them into a separate rating form (Appendix F)*. Each Test Administrator and Observer utilized a single scoresheet per Examinee per performance test. Upon completion of scoring an Examinee, the scoresheets were collected and inspected by USAIHRU personnel to insure that all performance measures were scored by all Test Administrators and Observers.

The prototype performance tests consisted of a variable number of performance measures upon which each Examinee was to be evaluated. Table 3 indicates the number of performance measures contained in each test.

The Test Administrator and the four Observers were instructed to rate each Examinee on each performance measure for a given performance test in terms of a "YES" or "NO" score. Their ratings were based upon the criteria

* Appendixes are available from ARI.

Table 3

NUMBER OF PERFORMANCE MEASURES PER PROTOTYPE PERFORMANCE TEST

Test Item	No. of PMs
Prepare M72A2 LAW for Firing; Restore M72A2 LAW to Carrying Configuration	12
Install/Recover an Electrically Armed Claymore Mine	8
Apply the Four Life-Saving Measures	18
Camouflage Yourself and Your Individual Weapon	11

contained in each performance measure and the evaluation instructions provided by the performance test. The score "YES" for any particular performance measure was to be given to an Examinee if he successfully and completely performed the action required of him according to the criteria. A "NO" score for a particular performance measure was to be given to any Examinee who failed to meet the standard established by the criteria. An overall "GO" or "NO GO" score was given to each Examinee as an indication of his total performance on each task. A "GO" score was based on the number of "YES" scores obtained, and the criteria provided in the test. For Test Items 1, 2, and 4, a "GO" score was given to an Examinee only if he successfully completed all of the individual performance measures within a task. For Test Item 3, a "GO" score was given to an Examinee if five or less "NO" scores were obtained, and the "NO" scores did not include specified critical performance measures.

Questionnaires

In order to determine the face validity of the performance tests, questionnaires were developed and administered to Examinees, Observers, and Test Administrators. The procedure for administering questionnaires was as follows:

1. Upon completion of testing on a specific performance test, each Examinee was administered the Examinee Questionnaire appropriate to that particular test. This procedure was followed for all four performance tests.

2. Upon completion of testing for all 15 Examinees on a specific performance test, the Test Administrator and the four Observers were administered the Test Administrator and Observer Questionnaires, respectively. This procedure was followed for all four performance tests.

Each questionnaire was developed with respect to the frame of reference of the individual responding. Anonymity and confidentiality of response was ensured. A brief description of the information that each questionnaire was designed to obtain is provided below. The complete questionnaires are provided in Appendix G.*

Examinee Questionnaire - designed to determine the Examinees' impressions and opinions toward the job relevance, degree of realism, and fairness of each of the four performance tests administered.

Observer Questionnaire - designed to determine the Observers' impressions and opinions toward the job relevance, degree of realism, ease of rating, and fairness of each of the four performance tests administered.

Test Administrator Questionnaire - divided into two parts. Part I was designed to determine the Test Administrators' impressions and opinions toward the job relevance, degree of realism, feasibility of administration, and fairness of each of the four performance tests administered. Part II was concerned with determining the amount and type of administrative preparation undertaken by the Test Administrator in preparing for the conduct of each of the four performance tests.

---

* Appendixes are available from ARI.

RESULTS AND DISCUSSION

VARIABILITY IN PERFORMANCE TEST SITUATIONS

One of the major issues associated with the implementation of hands-on performance testing during annual MOS evaluations is the extent to which standardized testing situations will be obtained across the Army. In order to provide initial information on the nature and extent of this potential problem, the first objective of the field investigation was concerned with test standardization in terms of test site preparation and test administration. Five NCOs with comparable backgrounds and experience were given the same instructions and materials, and the objective was to identify differences in the test sites that they established and differences in their administrative procedures. It was expected that information in these areas would assist in defining the training requirements for test administrators and identify additional content areas required in the development and writing of performance tests.

Variability in Test Site Preparation

In general, there were a large number of differences between the test sites established by the five Test Administrators. These differences were examined and screened to identify those factors that appeared to have an effect on the Examinee's evaluation, or at least had the potential of resulting in differential evaluations.

LAW Test. The site preparation section of this performance test specified that a ready rack should be cut in the front wall of the individual fighting position for storing the LAWs. Three of the five Test Administrators did not prepare the ready rack as indicated. Since the positioning of equipment was a critical aspect of this test, the absence of ready racks for storage may have influenced some of the evaluations.

This test involved the use of one defective LAW and one operational LAW. The weapons were supposed to be positioned in such a manner that the Examinee would pick up the defective LAW first. In addition to the problem of differences in test sites, there is also a problem related

to performance test development and the specification of individual performance measures for observation. In this case, the rater's observation of the first performance measure in the test depends on actions taken by the Examinee. The first performance measure of this test indicated that the soldiers should have discarded the defective launcher and picked up the second LAW. If the Examinee did not pick up the defective LAW first, either because of the Test Administrator's placement or other circumstances, he may have failed the test. The implication of this problem for test development is that the observation of individual performance measures for a test should not be contingent on the actions of the Examinee.

Camouflage Test. Since very little site preparation was required for this test, only a few differences between test sites were observed. One of these modifications probably had an effect on the Examinees' performance and subsequent evaluations. For two of the groups, the Test Administrator provided the Examinees with a mirror to assist them in applying camouflage materials to their skin. Eight of the performance measures in this test were concerned with the effectiveness of camouflaging the face and other areas of exposed skin. For the two groups that were provided a mirror, the percent of Examinees that received "Yes" ratings in either one or both of these groups was higher than the overall average on seven of the eight skin camouflaging performance measures.

It appears that the implication of this site preparation deviation for test development is that the equipment requirements and conditions should be more clearly specified. In this particular case, standardized testing situations might have been achieved if it had been clearly stated in the equipment section that only those items listed would be permitted in the testing situation.

Life-Saving Test. Several modifications were made in the life-saving test sites which probably had some impact on the ease with which certain performance measures were completed or the time required to perform the actions. Two of the Test Administrators positioned the manikin on tables,

while the Examinees in the other groups were required to complete the test with the manikin placed on the ground. Also, two of the Test Administrators did not attach the poncho, which was required for completion of some of the performance measures, to the manikin's load-bearing equipment. Both of these modifications in the test site probably had some influence on the time required to complete the test. Since there was a time standard associated with the performance in each test, there may have been some impact on the Examinees' overall evaluation.

For one of the groups, the Test Administrator provided the Examinees with an ammunition box for holding the manikin's feet after they had been elevated. The evaluation for the Examinees in this group were probably more precise than for the other groups. When the ammunition box was used, all judgment was, in effect, removed from those performance measures concerned with elevating the manikin's feet.

Claymore Test. The site preparation section of this test indicated that aiming stakes or markers should be placed five meters to the left, right, front, and rear of the aiming point for judging the accuracy of aiming the mine. Three of the Test Administrators did not use aiming stakes, which probably made the judgment of aiming accuracy much more difficult.

For all four of the performance tests, the differences observed in site preparation appeared to be due to insufficient detail and specification in the performance test or failure by the Test Administrators to follow instructions. Both of these problems could very likely be solved by increasing quality control of the test development process and providing training and more detailed instructions for the Test Administrators.

Variability in Administrative Procedures

The second major aspect of variability between test situations established by different Test Administrators was concerned with differences in administrative procedures used during the test. The degree of variability in test administration across all four performance tests is presented in Table 4. The data given in this table represent a summary of all 75

Table 4

VARIABILITY IN TEST ADMINISTRATION

Characteristics of Test Administration	Percent of Administrations Given "Yes" (Total N = 75 for Each Test)			
	LAW	Camouflage	Life Saving	Claymore
1. Instructions were read verbatim to the Examinee from the test scenario	75.7	79.2	73.5	43.3
2. All activities described in the test scenario were completed	61.6	56.9	60.3	71.6
3. The Test Administrator coached or provided feedback to the Examinee during the test	35.6	26.4	57.4	19.4

administrations for each test in terms of the percent of "Yes" checks given by the HumRRO observer.

On the average, the instructions were read verbatim to the Examinee 68 percent of the time. The percent for the Claymore test was considerably lower than the average, probably because this particular test involved the use of two different locations. In approximately 37 percent of the administrations, at least one deviation from the instructions or the scenario was observed. Although the Test Administrators were specifically instructed not to coach the Examinees or provide any type feedback, it appeared to be very difficult for the NCOs to divorce themselves from their normal role as a trainer. The instances of coaching were found to be much higher for the Life-Saving Test, which was probably due to the relatively low proficiency of the Examinees on first aid tasks. When coaching did occur, it took the form of verbal feedback 95 percent of the time.

A content analysis of the various deviations from the test scenario is shown in Table 5. These deviations by the Test Administrator did not include differences in test site preparation, but only those modifications that occurred during the actual administration of the test. For the category dealing with failure to restore the test site to the original condition between administrations, the following types of deviations were found:

(1) the components of the Claymore were not replaced in the bandoleer, (2) the bleeding leg wound was covered and not visible, (3) the load-bearing equipment was not replaced on the manikin, (4) used camouflage materials were not removed from the test site, and (5) the crack in the LAW was not always placed face down. An example of a change in the sequence of activities, which occurred 12 times, was the Test Administrator instructing the Examinees to inspect the LAW before they had an opportunity to initiate that action without being cued.

Table 5

CONTENT ANALYSIS OF DEVIATIONS FROM THE TEST SCENARIO BY THE TEST ADMINISTRATORS

Frequency	Type of Deviation
89	Instructions were not read verbatim from the test scenario
59	Equipment and materials at the test site were not restored to original conditions between administrations
30	Electrical circuit for the Claymore was not checked
27	Sight alignment of the Claymore was not checked
24	Camouflaged weapon was not checked for possible problems in sighting and functioning
23	Sections of the instructions to the Examinees were omitted
21	Sequence of activities in the test scenario was modified

RELIABILITY OF OBSERVATION IN PERFORMANCE TEST SITUATIONS

Another major issue which relates to performance test development, administration, and scoring is the degree of agreement between independent raters observing the same performance test. If the degree of agreement is low between independent raters, the instrument would not be reliable and could not be used for personnel evaluations. In this situation, it would be necessary to revise the test until the performance measures for evaluation

are observable and explicit, and also provide the raters with more detailed instructions and information on the behaviors being rated.

In the present study, each group consisting of the Test Administrator and four Observers rated the performance of 15 Examinees on each of the four tests. This provided five independent ratings of the same performance on each of 60 test administrations.

The statistic used to measure the degree of interrater reliability was the Cochran Q Test, which tested the null hypothesis that the "successes" and "failures" were randomly distributed in the rows and columns of the table. The Cochran Test was conducted for each combination of group, performance test, and performance measure. The values of Q and the level of significance are given in Tables 6-9. In addition, the percent of agreement between raters for each performance measure are presented in Appendix H.

Table 6

COCHRAN Q COEFFICIENTS FOR EACH PERFORMANCE MEASURE IN THE LAW TEST

Performance Measure	Group A	Group B	Group C	Group D	Group E
1	2.40	24.00***	0	4.00	8.70
2a	22.67***	32.30***	9.23	12.00*	18.51***
2b	6.00	19.18***	5.14	13.33**	4.53
2c	20.24***	15.16**	16.93**	1.60	3.08
3a	16.00**	4.67	8.89	8.00	0
3b	9.81*	3.33	18.13**	6.40	3.00
3c	12.00*	1.33	9.33	8.00	3.00
3d	5.40	11.20*	4.58	4.00	6.40
3e	6.28	32.80***	18.22**	5.60	6.67
4	12.44*	20.90***	6.61	6.86	0.57
5	5.60	7.83	0.80	10.00*	12.24*
6	2.00	15.09**	7.08	4.00	4.00

*<.05, df = 4
**<.01, df = 4
***<.001, df = 4

Table 7

COCHRAN Q COEFFICIENTS FOR EACH PERFORMANCE MEASURE IN THE CAMOUFLAGE TEST

Performance Measure	Group A	Group B	Group C	Group D	Group E
1a	10.00*	17.24**	5.45	6.40	3.06
1b	7.50	43.07***	8.62	4.00	2.13
1c	19.46***	8.15	9.41	4.00	1.09
1d	8.61	13.69**	7.78	4.00	6.25
1e	5.89	3.83	13.25*	5.00	5.30
1f	5.28	14.62**	15.47**	5.71	6.59
1g	10.37*	18.34**	11.52*	1.60	4.00
1h	9.89	30.34***	2.15	5.30	23.09***
2	21.74***	22.34***	12.67*	8.00	20.32***
3a	2.00	1.92	12.16*	4.00	20.00***
3b	9.50*	3.50	3.11	4.00	7.66

Table 8

COCHRAN Q COEFFICIENTS FOR EACH PERFORMANCE MEASURE IN THE LIFESAVING TEST

Performance Measure	Group A	Group B	Group C	Group D	Group E
1	6.00	4.00	4.57	3.11	1.71
2	1.33	2.44	0	20.50***	6.22
3a	4.00	4.00	3.00	2.20	12.53*
3b	65.33***	6.15	12.38*	12.00*	6.15
3c	4.24	7.04	0	7.20	3.00
3d	11.43*	1.27	20.67***	23.60***	15.79**
4a	1.33	3.00	8.00	2.00	3.05
4b	3.47	13.60**	7.05	13.23*	6.28
4c	1.60	3.00	3.00	4.00	5.00
5	7.08	5.30	3.50	4.00	14.43**
6a	8.31	21.14***	8.80	23.17***	17.71**
6b	15.75**	16.14**	15.29**	4.00	24.00***
6c	13.40**	23.20***	2.74	12.60*	18.35**
7a	2.29	4.31	5.74	34.75***	1.03
7b	10.46*	7.20	11.20*	9.33	9.45
7c	5.11	4.00	9.88*	8.50	15.33**
7d	8.15	6.28	11.43*	31.06***	13.33**
7e	22.72***	6.61	18.25**	34.05***	17.83**

*<.05, df = 4
**<.01, df = 4
***<.001, df = 4

Table 9

COCHRAN Q COEFFICIENTS FOR EACH PERFORMANCE MEASURE IN THE CLAYMORE TEST

Performance Measure	Group A	Group B	Group C	Group D	Group E
1	3.00	6.14	9.33	0	2.20
2	4.63	10.60*	11.20*	10.22*	8.75
3	21.28***	17.95**	3.56	6.40	5.80
4	12.88*	7.40	8.94	16.00**	2.86
5	11.77*	4.76	2.50	9.81*	2.35
6	4.00	14.88**	4.00	0	4.00
7	4.00	6.88	18.00**	6.40	4.00
8	14.30**	5.63	2.91	2.80	13.39**

*<.05, df = 4
**<.01, df = 4
***<.001, df = 4

Since there was a considerable amount of variability between groups for a given test, an arbitrary criterion was selected to summarize the data for purposes of discussion. For a given test and performance measure, the Q values were examined for all five groups. The performance measure was considered to have significantly low interrater reliability if two of the groups had significant Q values with one at the .01 level or three of the groups had significant Q values at the .05 level or less. When this criterion was applied to the four tables of Q values, a significant amount of disagreement between independent raters was obtained on the performance measures listed in Table 10.

Table 10

PERFORMANCE MEASURES WITH SIGNIFICANT DIFFERENCES BETWEEN INDEPENDENT RATERS

Performance Test	Performance Measure
LAW	2a, 2b, 2c 3e
Camouflage	1f, 1g, 1h 2
Life-Saving	3b, 3d 6a, 6b, 6c 7d, 7e
Claymore	2 3 8

Summarizing across all groups, the degree of agreement between raters on these performance measures was found to be quite low. The procedural checklists and field test notes were examined for possible sources of the disagreement between raters. The following situations were found which provide possible explanations for the low degree of agreement:

1. The evaluation of several performance measures was dependent on the Examinee's verbal report, which may have created a situation of low reliability because of different terminology or audibility.

2. Some performance measures required that the preceding actions be completed in a specific sequence, and it appeared that some raters strictly followed the criterion while others did not.

3. Some performance measures were ambiguous statements which were open to the interpretation and bias of the individual rater.

4. Some performance measures appeared to be interpreted differently as a function of specific unit SOPs.

5. Two separate actions were included in single performance measures, and it is not known whether the raters required one or both of the actions to be completed for a "Yes" score.

6. Several performance measures overlapped or were repeated later in the sequence which appeared to create confusion in scoring among the raters.

7. Sequences in the scenario which were timed were not clearly identified, and this appeared to result in different evaluations of whether or not the time standard had been met.

8. Some performance measures in the scenario were out of sequence with respect to the order in which they would be scored.

It appeared that the situations described above tended to result in a mix of "Yes" and "No" scores among the five raters. Situations that produce low interrater reliability must be modified and improved before the performance test can be used for evaluation purposes. It appeared that the large majority of these problems could be solved by improving the test development process and providing additional instructions and training for the raters.

FACE VALIDITY OF PERFORMANCE TESTS

A basic objective of this phase of the research was to determine if the performance tests were perceived to be job relevant, and also fair and reasonable tests of a soldier's ability to perform certain tasks. If the Examinee does not perceive the test to be relevant to his job and generally within his capability, various degrees of protest will likely result. This is not to suggest that the tests should be modified to coincide with perceptions, but that the Examinees should be given a clear and accurate picture of what is expected of them on the job and, therefore, in the test situation.

Test Administrators and Observers

Since many of the items on the questionnaires for the Test Administrators and the Observers were identical, their responses were combined into a group of 25 NCOs. With respect to the job relevance of the performance tests, 95 percent of the NCOs felt that the tests were a realistic sample of the Examinees' job duties. In terms of the perceived fairness and equity of the tests, 95 percent of the NCOs believed that all of the Examinees had an equal chance of passing the tests. The preceding percentages were obtained by summarizing over all four tests, and it should be pointed out that the LAW and Lifesaving tests were slightly lower on both job relevance and fairness than the other two tests. Almost all the NCOs (98 percent) felt that the average combat soldier could pass the tests if he were given the necessary study time and materials.

Approximately 90 percent of the NCOs believed that the tests would be practical and feasible for administration in a unit environment or an FTX/ATT situation. With the possible exception of the Lifesaving test, the performance tests were considered to be very easy to administer and/or

rate Examinee performance. Considering only the five Test Administrators, an average of 12 hours was devoted to preparing and planning for the administration of each test. Depending on the particular test involved, one to three of the Test Administrators felt the need to seek assistance in conducting the performance test.

Examinee Performance

The overall test performance of the 75 Examinees on all four tests is shown in Table 11. In terms of the overall test scores, the percent of Examinees passing a given test was quite low. The low passing rates were

Table 11

PERCENT OF EXAMINEES RECEIVING "GO" RATINGS FROM THE TEST ADMINISTRATOR ON EACH TEST

Performance Test	Group A	Group B	Group C	Group D	Group E	Overall
LAW	46.7 (n=15)	0 (n=15)	0 (n=15)	73.3 (n=15)	14.3 (n=14)	27.0 (n=74)
Camouflage	26.7 (n=15)	0 (n=15)	0 (n=13)	71.4 (n=14)	33.3 (n=15)	26.4 (n=72)
Lifesaving	13.3 (n=15)	0 (n=15)	0 (n=13)	42.9 (n=14)	0 (n=12)	11.6 (n=69)
Claymore	0 (n=15)	6.7 (n=15)	18.2 (n=11)	30.8 (n=13)	50.0 (n=14)	20.6 (n=68)

probably due to the criterion established for passing the tests. All of the tests except Lifesaving required a "Yes" on all performance measures in order to receive a "Go" for the test. Although up to five performance measures in the Lifesaving test could be failed, eight specific performance measures could not be missed.

The proficiency level of the Examinees was actually much higher than indicated by the overall test scores. The percent of Examinees receiving "Yes" ratings from the Test Administrator on each performance measure of each test are given in Tables 12-15. Although there was a considerable amount of variability across tests and groups, the majority of the passing rates were in the 65 to 85 percent range.

Table 12

PERCENT OF EXAMINEES RECEIVING "YES" RATINGS FROM THE TEST ADMINISTRATOR ON THE LAW TEST

Performance Measure	Group A	Group B	Group C	Group D	Group E	Overall
1	35.7 (n=14)	13.3 (n=15)	13.3 (n=15)	73.3 (n=15)	42.9 (n=14)	35.6 (n=73)
2a	26.7 (n=15)	13.3 (n=15)	35.7 (n=14)	93.3 (n=15)	100 (n=14)	53.4 (n=73)
2b	60.0 (n=15)	73.3 (n=15)	80.0 (n=15)	93.3 (n=15)	92.9 (n=14)	79.7 (n=74)
2c	33.3 (n=15)	84.6 (n=13)	21.4 (n=14)	93.3 (n=15)	100 (n=14)	66.2 (n=71)
3a	100 (n=14)	93.3 (n=15)	73.3 (n=15)	100 (n=15)	100 (n=14)	97.3 (n=73)
3b	100 (n=15)	73.3 (n=15)	93.3 (n=15)	100 (n=15)	100 (n=14)	93.2 (n=74)
3c	100 (n=15)	93.3 (n=15)	86.7 (n=15)	100 (n=15)	100 (n=14)	96.0 (n=74)
3d	66.7 (n=15)	57.1 (n=15)	66.7 (n=15)	93.3 (n=15)	85.7 (n=14)	73.0 (n=74)
3e	83.3 (n=12)	80.0 (n=15)	42.9 (n=14)	93.3 (n=15)	78.6 (n=14)	75.7 (n=70)
4	80.0 (n=15)	64.3 (n=14)	60.0 (n=15)	86.7 (n=15)	64.3 (n=14)	71.2 (n=73)
5	60.0 (n=15)	40.0 (n=15)	60.0 (n=15)	86.7 (n=15)	57.1 (n=14)	60.8 (n=74)
6	73.3 (n=15)	66.7 (n=15)	53.3 (n=15)	93.3 (n=15)	100 (n=14)	77.0 (n=74)

Table 13

PERCENT OF EXAMINEES RECEIVING "YES" RATINGS FROM THE TEST ADMINISTRATOR ON THE CAMOUFLAGE TEST

Performance Measure	Group A	Group B	Group C	Group D	Group E	Overall
1a	93.3 (n=15)	40.0 (n=15)	61.5 (n=13)	78.6 (n=14)	80.0 (n=15)	70.8 (n=72)
1b	93.3 (n=15)	46.7 (n=15)	69.2 (n=13)	100 (n=14)	86.7 (n=15)	79.2 (n=72)
1c	100 (n=15)	53.3 (n=15)	61.5 (n=13)	100 (n=14)	84.6 (n=13)	80.0 (n=70)
1d	93.3 (n=15)	60.0 (n=15)	69.2 (n=13)	100 (n=14)	73.3 (n=15)	79.2 (n=72)
1e	86.7 (n=15)	35.7 (n=14)	41.7 (n=12)	85.7 (n=14)	60.0 (n=15)	62.9 (n=70)
1f	86.7 (n=15)	53.3 (n=15)	23.1 (n=13)	92.9 (n=14)	71.4 (n=14)	66.2 (n=71)
1g	93.3 (n=15)	71.4 (n=14)	38.5 (n=13)	92.9 (n=14)	78.6 (n=14)	75.7 (n=70)
1h	53.3 (n=15)	73.3 (n=15)	69.2 (n=13)	100 (n=14)	60.0 (n=15)	70.8 (n=72)
2	73.3 (n=15)	86.7 (n=15)	84.6 (n=13)	100 (n=14)	53.3 (n=15)	79.2 (n=72)
3a	73.3 (n=15)	60.0 (n=15)	84.6 (n=13)	100 (n=14)	86.7 (n=15)	80.6 (n=72)
3b	66.7 (n=15)	86.7 (n=15)	84.6 (n=13)	92.9 (n=14)	86.7 (n=15)	83.3 (n=72)

Table 14

PERCENT OF EXAMINEES RECEIVING "YES" RATINGS FROM THE TEST ADMINISTRATOR ON THE LIFESAVING TEST

Performance Measure	Group A	Group B	Group C	Group D	Group E	Overall
1	93.3 (n=15)	40.0 (n=15)	50.0 (n=12)	85.7 (n=14)	75.0 (n=12)	69.1 (n=68)
2	54.5 (n=11)	0 (n=15)	25.0 (n=12)	64.3 (n=14)	25.0 (n=12)	32.8 (n=64)
3a	92.9 (n=14)	100 (n=15)	91.7 (n=12)	78.6 (n=14)	100 (n=12)	92.5 (n=67)
3b	66.7 (n=6)	58.3 (n=16)	66.7 (n=12)	78.6 (n=14)	36.4 (n=11)	61.8 (n=55)
3c	85.7 (n=14)	80.0 (n=15)	100 (n=12)	85.7 (n=14)	66.7 (n=12)	83.6 (n=67)
3d	64.3 (n=14)	53.3 (n=15)	50.0 (n=12)	78.6 (n=14)	75.0 (n=12)	64.2 (n=67)
4a	73.3 (n=15)	80.0 (n=15)	41.7 (n=12)	71.4 (n=14)	50.0 (n=12)	64.7 (n=68)
4b	78.6 (n=14)	28.6 (n=14)	33.3 (n=12)	85.7 (n=14)	66.7 (n=12)	59.1 (n=66)
4c	76.9 (n=13)	57.1 (n=14)	33.3 (n=12)	92.9 (n=14)	33.3 (n=12)	60.0 (n=65)
5	92.9 (n=14)	86.7 (n=15)	91.7 (n=12)	92.9 (n=14)	54.5 (n=11)	84.8 (n=66)
6a	36.4 (n=11)	20.0 (n=15)	25.0 (n=12)	71.4 (n=14)	8.3 (n=12)	32.8 (n=64)
6b	78.6 (n=14)	57.1 (n=14)	36.4 (n=11)	92.9 (n=14)	66.7 (n=12)	67.7 (n=65)
6c	73.3 (n=15)	40.0 (n=15)	25.0 (n=12)	78.6 (n=14)	33.3 (n=12)	51.5 (n=68)

(continued)

Table 14 (continued)

Performance Measure	Group A	Group B	Group C	Group D	Group E	Overall
7a	86.7 (n=15)	46.7 (n=15)	75.0 (n=12)	92.9 (n=14)	58.3 (n=12)	72.1 (n=68)
7b	80.0 (n=15)	13.3 (n=15)	58.3 (n=12)	100 (n=14)	50.0 (n=12)	60.3 (n=68)
7c	71.4 (n=14)	80.0 (n=15)	58.3 (n=12)	92.9 (n=14)	91.7 (n=12)	79.1 (n=67)
7d	33.3 (n=15)	20.0 (n=15)	25.0 (n=12)	85.7 (n=14)	45.4 (n=11)	41.8 (n=67)
7e	53.3 (n=15)	53.3 (n=15)	25.0 (n=12)	78.6 (n=14)	58.3 (n=12)	54.4 (n=68)

Table 15

PERCENT OF EXAMINEES RECEIVING "YES" RATINGS FROM THE TEST ADMINISTRATOR ON THE CLAYMORE TEST

Performance Measure	Group A	Group B	Group C	Group D	Group E	Overall
1	93.3 (n=15)	57.1 (n=14)	63.6 (n=11)	100 (n=13)	85.7 (n=14)	80.6 (n=67)
2	28.6 (n=14)	21.4 (n=14)	54.5 (n=11)	76.9 (n=13)	85.7 (n=14)	53.0 (n=66)
3	66.6 (n=15)	71.4 (n=14)	81.8 (n=11)	84.6 (n=13)	85.7 (n=14)	77.6 (n=67)
4	35.7 (n=14)	21.4 (n=14)	81.8 (n=11)	61.5 (n=13)	71.4 (n=14)	53.0 (n=66)
5	78.6 (n=14)	35.7 (n=14)	45.4 (n=11)	61.5 (n=13)	85.7 (n=14)	62.1 (n=66)
6	100 (n=15)	64.2 (n=14)	100 (n=11)	100 (n=13)	100 (n=13)	92.4 (n=66)
7	93.3 (n=15)	85.7 (n=14)	100 (n=10)	84.6 (n=13)	92.8 (n=14)	90.9 (n=66)
8	33.3 (n=15)	42.8 (n=14)	72.7 (n=11)	84.6 (n=13)	92.8 (n=14)	64.2 (n=67)

Examinees

The 75 Examinees were requested to rate the importance of the task covered in each performance test with respect to the performance of their MOS duties. On a five-point rating scale (5 = Extremely Important), the mean ratings ranged from 3.99 to 4.43 which indicated that the tasks were judged to be between highly important and extremely important in the performance of MOS-related duties. Approximately 93 percent of the Examinees believed that their chances were equal or better for passing the test as compared with any other soldier with the same MOS and skill level. When the Examinees were asked if they felt they had passed the test, substantial discrepancies were obtained between their perceptions and the actual test results, as shown in Table 16.

Table 16

PERCEIVED AND ACTUAL TEST RESULTS

Performance Test	Percent of Examinees that Believed they Passed the Test	Percent of Examinees Receiving "GO" Ratings from the Test Administrator
LAW	82.2	27.0
Camouflage	95.7	26.4
Lifesaving	64.7	11.6
Claymore	66.7	20.6

In the present field test, the Examinees were given very little information concerning the performances required in the test situation. The discrepancies shown in Table 16, however, suggest that the Examinees should be given detailed information concerning what will be expected of them in the test situation, to include scoring criteria. If information of this nature is not provided to the Examinee well in advance of the test, it is not likely the test will be perceived as fair and reasonable.

IMPLICATIONS FOR EPMS

The overall objective of the field investigation was completely satisfied in that numerous problem areas were identified with various

degrees of quantification. The implications of these findings for EPMS and the implementation of hands-on performance testing in MOS evaluations appeared to be related to three major areas.

Feasibility Testing

Although the performance tests developed by the Infantry School were not produced under ideal conditions, the level of effort devoted to the four tests was probably fairly high compared to that which may be expected under full EPMS implementation. Regardless of the level of effort during test development and the degree of input from subject matter experts, the results of this study suggest that each performance test developed as a candidate for MOS evaluation should be thoroughly field tested, revised, and tested again before implementation.

Performance Test Development Process

The research completed in this phase of the project indicated that at least two areas in the test development process should be given greater emphasis and more detailed specification. The first area is concerned with the description of administrative instructions and procedures for the Test Administrator. In the field test, the degree of test standardization was reduced considerably through deviations or modifications made by the Test Administrators in the areas of test site preparation and test administration. It is believed that the extent of test standardization could be substantially increased by increasing the amount and precision of information provided to the Test Administrator. The second area of test development which should be improved is the identification and specification of performance measures which are to be scored. Many performance measures were identified that were ambiguous statements, included two separate actions, required verbal reports, or were not in the appropriate sequence. All of these problems tended to reduce the degree of interrater reliability. It appeared that most of these problems associated with the specification of performance measures could be eliminated by refining the test development process.

Test Administrator Training

The results of the field test indicated that, in order to achieve standardized testing situations and reliability of performance measurement, considerably more effort will be necessary in the selection and training of Test Administrators. Numerous examples were found of failure to follow instructions and introduction of individual interpretation and bias. It appears that the Test Administrators probably should be subject matter experts in the task areas included in the tests. In addition to being given a thorough orientation on performance testing situations and the implications of various types of procedural errors, the Test Administrators probably should test 20-30 subjects on each test for which they have responsibility. Until the Test Administrators experience most of the possible Examinee reactions and problems, they cannot react to these various situations in any standardized manner.