Technical Report 1128

DEVELOPMENT OF PREDICTOR AND CRITERION MEASURES FOR THE NCO21 RESEARCH PROGRAM

Deirdre J. Knapp
Jennifer L. Burnfield
Chris E. Sager
Gordon W. Waugh
John P. Campbell
Charlie L. Reeve
Roy C. Campbell
Human Resources Research Organization

Leonard A. White
Tonia S. Heffner
U.S. Army Research Institute

June 2002

United States Army Research Institute
for the Behavioral and Social Sciences

U.S. Army Research Institute for the Behavioral and Social Sciences

A Directorate of the U.S. Total Army Personnel Command

ZITA M. SIMUTIS
Acting Director

Research accomplished under contract
for the Department of the Army

Human Resources Research Organization

Technical Review by

Lynn M. Milan
Lisa J. Mills

NOTICES

DISTRIBUTION: Primary distribution of this Technical Report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, Attn: TAPC-ARI-PO, 5001 Eisenhower Ave., Alexandria, VA 22333-5600.

FINAL DISPOSITION: This Technical Report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

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

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy)	2. REPORT TYPE	3. DATES COVERED (from. . . to)
June 2002	Interim	October 1999 – February 2001

4. TITLE AND SUBTITLE	5a. CONTRACT OR GRANT NUMBER
Development of Predictor and Criterion Measures for the NCO21 Research Program	DASW01-98-D-0047, DO 0015
	5b. PROGRAM ELEMENT NUMBER 622785
6. AUTHOR(S) Knapp, Deirdre J., Burnfield, Jennifer L., Sager, Christopher E., Waugh, Gordon W., Campbell, John P., Reeve, Charlie L. Campbell, Roy C. (HumRRO), White, Leonard A., Heffner, Tonia S. (ARI)	5c. PROJECT NUMBER A790
	5d. TASK NUMBER 1218
	5e. WORK UNIT NUMBER C05

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Human Resources Research Organization 66 Canal Center Plaza, Suite 400 Alexandria, VA 22314	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. MONITOR ACRONYM
U. S. Army Research Institute for the Behavioral & Social Sciences 5001 Eisenhower Avenue Alexandria, VA 22333	ARI
	11. MONITOR REPORT NUMBER Technical Report 1128

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

Tonia Heffner, Contracting Officer's Representative

14. ABSTRACT *(Maximum 200 words)*:

The NCO21 research program was undertaken to help the U.S. Army plan for the impact of future demands on the noncommissioned officer (NCO) corps. The performance requirements and associated knowledge, skills, and aptitudes (KSAs) expected of future successful NCOs were used as a basis for developing tools that could be incorporated into an NCO performance management system geared to 21st-century job demands. This report documents the design and development of predictor and criterion measures that will be used in a criterion-related validation data collection.

The predictor measures include the Armed Services Vocational Aptitude Battery (ASVAB), Assessment of Individual Motivation (AIM), and Biographical Information Questionnaire (BIQ), which are operational tests already used in the Army for other purposes. A written Situational Judgment Test (SJT), the Experience and Activities Record (ExAct), Personnel File Form (PFF21), and a semi-structured interview were developed for this project.

Two types of rating scale instruments were developed for gathering criterion data. The Observed Performance Rating Scales ask supervisors to rate soldiers on how well they perform in their current jobs. The Expected Future Performance Rating Scales ask supervisors to predict how their soldiers would perform in specific sets of conditions expected to be characteristic of future Army requirements.

15. SUBJECT TERMS

Behavioral and social science — Personnel
Selection and Classification — Manpower

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON (Name and Telephone Number)
16. REPORT	17. ABSTRACT	18. THIS PAGE			
Unclassified	Unclassified	Unclassified	Unlimited		Tonia Heffner (703) 617-8557 DSN 767-8557

Standard Form 298

Technical Report 1128

DEVELOPMENT OF PREDICTOR AND CRITERION MEASURES FOR THE NCO21 RESEARCH PROGRAM

Deirdre J. Knapp
Jennifer L. Burnfield
Chris E. Sager
Gordon W. Waugh
John P. Campbell
Charlie L. Reeve
Roy C. Campbell
Human Resources Research Organization

Leonard A. White
Tonia S. Heffner
U.S. Army Research Institute

Selection and Assignment Research Unit
Trueman R. Tremble, Jr., Acting Chief

U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

June 2002

Army Project Number
2O262785A790

Personnel Performance and
Training Technology

FOREWORD

This project, entitled "NCO21: 21st-century Noncommissioned Officer Requirements," is being conducted by the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) under the sponsorship of the Office of the Deputy Chief of Staff for Personnel (ODCSPER). The goal of NCO21 is to conduct an analysis of future conditions and future job demands in order to identify critical performance predictors, or knowledges, skills, and aptitudes (KSAs), that may eventually be used to select and grow future noncommissioned officers (NCOs). This project has been divided into three phases. Completion of the first two phases was documented in earlier reports. Phase I was the development of a detailed research plan for identifying characteristics required of future NCOs. In Phase II, the methodological steps of the Phase I research plan were executed. Anticipated job requirements of 21st-century NCOs (for the years 2000 through 2025) were forecasted and the most important KSAs needed for success in Army jobs were estimated.

Phase III involves the remainder of the project activities, including development and validation of KSA measures. This report documents the first stage of Phase III, including the design and development, to include field testing, of predictor and criterion measures to be used in a forthcoming validation data collection. The information presented in this report was briefed to the Chief, Enlisted Division, Directorate of Military Personnel Management, Deputy Chief of Staff for Personnel (DCSPER) and the DCSPER Sergeant Major on 13 August 2001. It was briefed to U.S. Army Training and Doctrine Command (TRADOC) representatives on 11 October 2001. Uses of the tools developed in this effort will be determined in discussions with ODCSPER and TRADOC representatives based on the findings obtained from the Phase III validation.

The goal of the Selection and Assignment Research Unit of ARI is to conduct research, studies, and analysis on the measurement of aptitudes and performance of individuals to improve the Army's selection and classification, promotion, and reassignment of officers and enlisted soldiers. This research will provide the foundation for recommended improved promotion and development procedures for enlisted personnel.

MICHAEL G. RUMSEY
ActingTechnical Director

Acknowledgements

Many people contributed to the work documented in this report. Those not listed as authors include Dr. Michael G. Rumsey, Dr. Robert Kilcullen, and Ms. Emma Gregory from the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) and Dr. Laura Ford, Ms. Ani S. DiFazio, and Dr. Rodney A. McCloy from the Human Resources Research Organization (HumRRO). We gratefully acknowledge their ideas, feedback, and efforts related to this work. We also appreciate the tremendous secretarial support provided by Ms. LaVonda Murray of HumRRO.

DEVELOPMENT OF PREDICTOR AND CRITERION MEASURES FOR THE NCO21 RESEARCH PROGRAM

EXECUTIVE SUMMARY

Research Requirement

The NCO21 research program was undertaken to help the U.S. Army plan for the impact of future demands on the noncommissioned officer (NCO) corps. When the NCO21 research program began, a great deal of effort was being devoted to analyzing national and global trends (e.g., more complex technology with increasingly sophisticated capabilities, demographic changes) that would presumably affect the U.S. military in terms of its missions, organizational structure and technology, strategies and tactics, and personnel systems. But these analyses and forecasts were not available in any consolidated form. Indeed, there was (and still is) considerable variation in the prognostications being made. Moreover, very little had been done to look at the implications of expected future changes for the performance requirements of individual soldiers. The purpose of the first stage of this research program, then, was to (a) identify and review the available information on predictions and plans related to the Army's future and (b) attempt to abstract from these a reasonable idea of what performance expectations would be imposed on NCOs of the future. In subsequent stages of the research program, these expectations have been used to develop procedures and methods that could be incorporated into the NCO performance management system in an effort to make the NCO corps better prepared to handle 21st-century job demands. The purpose of this report is to document the design and development of predictor and criterion (job performance) measures that will be used in a criterion-related validation data collection. The report is primarily targeted toward a technical audience interested in the psychometric characteristics and quality of the measures.

Procedure

The NCO21 project team identified measurement methods that could be used to assess the broadest range of critical knowledges, skills, and aptitudes (KSAs) applicable across two eras (2000-2010 and 2010-2025). The team also identified measurement methods that could be used to assess NCO job performance. With the assistance of Army subject matter experts, the research team developed instruments and field tested them on approximately 500 soldiers. The instruments were finalized based on data analysis and lessons learned in the field test.

Findings

The project team identified seven predictor measures for use in the NCO21 project. Three measures—the Armed Services Vocational Aptitude Battery (ASVAB), Assessment of Individual Motivation (AIM), and Biographical Information Questionnaire (BIQ)—are operational tests already used in the Army for other purposes. Experimental versions of the AIM and BIQ were prepared for use in the present research. Four measures—a written Situational Judgment Test (SJT) (and its close cousin, the SJT-X), the Experience and Activities Record (ExAct), Personnel File Form (PFF21), and a semi-structured interview—were developed for

this project. Most of these instruments, however, made use of relevant, previously developed materials and items.

Two types of supervisor rating scale instruments were developed for gathering job performance criterion data. The Observed Performance Rating Scales ask supervisors to rate soldiers on how well they perform in their current jobs. The Expected Future Performance Rating Scales ask supervisors to predict how their soldiers would perform in specific sets of conditions expected to be characteristic of future Army requirements.

Utilization of Findings

Plans are to collect validation data from approximately 2,400 E4, E5, and E6 soldiers at seven Army installations from April through August 2001. The goal is to collect complete predictor data for E4 soldiers, complete predictor and criterion data for E5 soldiers, and partial predictor data (all except the interview) and complete criterion data for E6 soldiers. The primary purpose of the validation effort will be to determine what combination of KSA measures (i.e., performance predictors) best predicts important aspects of NCO performance (i.e., performance criteria). Based on the results, recommendations for further work supporting implementation of the most promising measures will be made.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

List of Tables

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

List of Figures

DEVELOPMENT OF PREDICTOR AND CRITERION MEASURES FOR THE NCO21 RESEARCH PROGRAM

CHAPTER 1: INTRODUCTION

This report describes the design and development of a set of predictor and criterion measures to be used in a criterion-related validation data collection to evaluate experimental noncommissioned officer (NCO) promotion procedures. This project is being conducted as part of a multi-phased research program sponsored by the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI).

Background

Overview of the NCO21 Research Program

The NCO21 research program was undertaken to help the U.S. Army understand and plan for the impact of future performance demands on the future NCO performance management system. A great deal of effort was being devoted to analyzing national and global trends (e.g., more complex technology with increasingly sophisticated capabilities, demographic changes) that will presumably affect the U.S. military in terms of its missions, organizational structure and technology, strategies and tactics, and personnel systems. But these analyses and forecasts were not available in any consolidated form. Indeed, there was (and still is) considerable variation in the prognostications being made. Moreover, very little had been done to look at the implications of expected future changes for the performance requirements of individual soldiers. The purpose of the first stage of this research program, then, was to (a) identify and review the available information on predictions and plans related to the Army's future and (b) attempt to abstract from these a reasonable idea of what performance expectations would be imposed on NCOs of the future. In subsequent stages of the research program, these expectations have been used to develop procedures and methods that could be incorporated into the NCO performance management system in an effort to make the NCO corps better prepared to handle 21st-century job demands.

Following some preliminary efforts conducted by ARI staff, the NCO21 research program was divided into three phases, each of which has been supported through a contract to the Human Resources Research Organization (HumRRO):

- Phase I: Develop a methodology to identify future job requirements. (Completed September, 1998.)
- Phase II: Forecast future NCO performance requirements and the individual characteristics necessary to meet those requirements. (Completed October, 1999.)
- Phase III: Develop measures of the relevant variables, conduct validation research to estimate their usefulness, and make recommendations for potential changes to the NCO promotion system. (Development of the measures is the subject of this report, the validation will be conducted in 2001, and the recommendations will follow).

The Phase II final report documents the collection and integration of future projections (Ford, R. Campbell, J. Campbell, Knapp, & Walker, 2000). The Phase II report also describes the construction of baseline (1990s) information about NCO requirements – both in terms of performance requirements (e.g., motivating and leading others) and in terms of the knowledges, skills, and aptitudes (KSAs) required for successful job performance (e.g., general cognitive aptitude, conscientiousness). The baseline requirements were then updated based on the analysis of conditions in two future eras (the period 2000-2010 and the period 2010-2025). Two expert panels (one comprising Army subject matter experts [SMEs] and another comprising personnel psychologists) used this information to judge the relative importance of KSAs for the different time periods. The products of Phase II thus included:

- Descriptions of the forecasted job demands for two future eras (2000-2010, 2010-2025).
- Lists of performance requirements for three eras (1990s baseline, 2000-2010, 2010-2025).
- Prioritized lists of KSAs for all three eras.

Because of differences in NCO requirements across ranks, the baseline and 2000-2010 era KSA priority rankings were determined separately by NCO level: junior (Corporal E4/E5), mid-level (E6/E7), and senior (E8/E9). The 2010-2025 era was forecast to incorporate the Army envisioned for the 2000-2010 era supplemented by a "Battleforce" component comprising more experienced and specialized soldiers. Therefore, the 2010-2025 era KSAs were prioritized simply for Battleforce NCOs, irrespective of rank.

In Phase III, the NCO21 project team has identified measurement methods that could be used to assess the broadest range of the most critical KSAs across the two future eras. The team has also identified measurement methods that could be used to assess NCO job performance. Following development and field testing of the required measures, these instruments will be used in a criterion-related validation data collection in 2001. The primary purpose of the validation effort will be to determine what combination of KSA measures (i.e., performance predictors) best predicts important aspects of NCO performance (i.e., performance criteria).

Whereas Phase II focused on soldier requirements across all NCO levels (shown in Table 1.1), the focus in Phase III has been narrowed to the semi-centralized NCO promotion system. This system covers promotions from grade E4 to E5 and from grade E5 to E6. It was necessary to narrow the focus because of the inordinate resources required to develop and validate measures suitable across all NCO ranks. The semi-centralized promotion system, however, covers more than 70% of the Army NCO corps, so improving this system would have a substantial impact.

Purpose of Report

The purpose of this report is to describe the design and development of the NCO21 predictor (KSA) and criterion (job performance) measures that will be used in the criterion-related validation data collection. It is primarily targeted toward a technical audience interested in the psychometric characteristics and quality of the measures.

Table 1.1. U.S. Army NCO Pay Grades and Ranks

Pay Grade	Rank
E4	Specialist or Corporal[a]
E5	Sergeant
E6	Staff Sergeant
E7	Sergeant First Class
E8	Master Sergeant
E9	Sergeant Major

[a]Most soldiers at the E4 level are Specialists; however, a small number are Corporals. Specialists are not NCOs; Corporals are considered junior NCOs.

Although the report chapters contain some limited discussions of implementation-related issues, this is not the focus of the present report. Ideas and specific recommendations for implementation will be discussed in-depth in a subsequent report. Those recommendations will be based on results of the validation research, reactions to the instruments by soldiers in the field, and input from Army stakeholders. The suggestions will, we hope, help address the complicated myriad of factors related to making a change to the Army's promotion processes (e.g., resource constraints, high volume of personnel actions).

The remainder of this chapter describes how the project team determined what measures to adopt, adapt, or develop for the purposes of this research. It also provides an overview of the development strategy, including pilot- and field-testing activities.

Determination of Measurement Instruments/Methods

Measurement Objectives

The primary objective was to identify or develop measurement procedures that would enable the project team to assess soldiers as thoroughly as possible on the relevant predictors (i.e., KSAs) and criteria (i.e., performance requirements) in the validation data collection. Additionally, the selected predictor measures were to be suitable for potential inclusion in the Army's operational semi-centralized NCO promotion system.

Table 1.2 illustrates the Phase III measurement objectives as they apply to predictor and criterion measurement. The predictor measures are designed for E4 and E5 soldiers (i.e., those soldiers eligible for promotion to E5 and E6) and the criterion measures are designed for evaluating the performance of E5 and E6 soldiers. The predictor measures assess E4 and E5 soldiers on KSAs associated with successful performance at the next higher grade (E5 and E6). Because this is a promotion system, relevant KSAs include how well the soldier performs in the current grade (e.g., their level of Military Occupational Specialty [MOS]-specific job knowledge), as well as more generic aptitudes and skills (e.g., general cognitive aptitude, spatial aptitude). Therefore, the lists of KSAs and performance requirements identified in Phase II overlap with each other.

Table 1.2. Phase III Measurement Goals

	Predictors	Criteria
Purpose	Instrument(s) that could be used to determine E4 and/or E5 soldiers' promotion potential	Instrument(s) that assess how well E5 and E6 soldiers perform their jobs
Target Grade	E4 and E5 soldiers	E5 and E6 soldiers
Target of Measurement	Knowledges, skills, and aptitudes (KSAs) relevant for promotion to E5 and E6 levels	E5 and E6 level performance requirements
Related Notes	Instruments must have potential for operational use. Relevant KSAs include performance in the soldier's current job, so the KSA and performance requirement lists overlap.	Instruments are intended for research use only so there is no need to be concerned about feasibility of operational use.

Predictor Measurement

The NCO21 KSAs identified in Phase II are listed and defined in Table 1.3.[1] The Phase II SMEs also provided judgments regarding the relative importance of the KSAs for current and future time periods. Although all KSAs in the list can be viewed as relevant, these judgments were used to help determine the KSAs that were most critical to measure in the NCO21 validation research effort.

Criterion Measurement

Phase II of the NCO21 project did not attempt to delineate specific task requirements for future NCOs, nor did it attempt to differentiate explicitly among performance requirements across NCO grades and time periods. It was simply not possible to abstract such specific predictions from the aggregate discussions and forecasts pertaining to the future Army. It did, however, identify a set of forecasted future NCO performance requirements. While still substantive in nature these expected future requirements were defined more generally than specific task responsibilities, which cannot be forecasted with any degree of certainty. Descriptions of the sets of future performance requirements and the procedures by which they were generated are described in the Phase II report. Because performance at the E4 and E5 levels can be used to evaluate promotion potential, these performance requirements are included in the KSA set listed in Table 1.3 (see items 12-38).

[1] Following Phase II, additional work was done on these KSAs to clarify each and distinguish among them. Thus this listing is slightly different than that provided in Ford et al. (2000).

Table 1.3. NCO21 Knowledges, Skills, and Aptitudes (KSAs) and Performance Requirements

Items 1-11 can be viewed as KSAs (i.e., predictors) only.

1. **Conscientiousness/Dependability**. The general tendency to be trustworthy, reliable, planful, and accountable. A general willingness to accept responsibility.
2. **General Cognitive Aptitude.** Has the overall capacity to understand and interpret information that is being presented, the ability to identify problems and reason abstractly, and the capability to learn new things quickly and efficiently.
3. **Need for Achievement.** Is generally predisposed to have confidence in own abilities and to seek and enjoy positions of leadership and influence. Would typically demonstrate enthusiasm and energy, and strive for accomplishment and recognition in almost any situation.
4. **Emotional Stability.** Has the tendency to act rationally and to display a generally calm, even mood. Typically maintains composure and is not overly distraught by stressful situations.
5. **Working Memory.** Has the ability to maintain information in memory for short periods of time and to retrieve it accurately.
6. **Spatial Relations Aptitude.** Has the ability to mentally visualize the relative positions of objects in two-dimensional or three-dimensional space, and how they will be positioned if they are moved or rotated in different ways.
7. **Perceptual Speed and Accuracy.** Has the ability to recognize and interpret visual information quickly and accurately, particularly with regard to comparing similarities and differences among words, numbers, objects, or patterns, when presented simultaneously or one after the other.
8. **Psychomotor Aptitude.** Has the ability to coordinate the simultaneous movements of one's limbs (arms, legs), to operate single controls or to operate multiple controls simultaneously, and to make precise control adjustments that involve eye-hand coordination.
9. **Basic Math Facility.** Knows and applies addition, subtraction, multiplication, division, and simple mathematical formulas.
10. **Basic Electronics Knowledge.** Knows general information regarding electronic principles and electronics equipment operation and repair. Knows general facts and principles relevant for a wide variety of electronics related tasks, but does not necessarily have highly specific electronics knowledge required for a particular job.
11. **Basic Mechanical Knowledge.** Knows general information regarding mechanical principles, tools, and mechanical equipment operation and repair. Knows general facts and principles relevant for a wide variety of tasks that require technical knowledge, but does not necessarily have highly specific mechanical knowledge required for a particular job.

Table 1.3. NCO21 Knowledges, Skills, and Aptitudes (KSAs) and Performance Requirements (Continued)

The remaining items can be viewed as either KSAs (predictors) or performance requirements (criteria).

12. **Problem-Solving/Decision Making Skill.** Reacts to new problem situations by applying previous experience and previous education/training appropriately and effectively. Does not apply rules or strategies blindly. Assesses costs and benefits of alternative solutions and makes timely decisions even with incomplete information.

13. **Writing Skill.** Communicates thoughts, ideas, and information successfully to others through writing. Uses proper sentence structure including grammar, spelling, capitalization, and punctuation.

14. **Oral Communication Skill.** Speaks in a clear, organized, and logical manner. Communicates detailed information, instructions, or questions in an efficient and understandable way. Note that this skill refers to how well the individual can speak and communicate, not whether technical expertise is high or low.

15. **MOS/Occupation-Specific Knowledge and Skill.** Possesses the necessary technical knowledge and skill to perform MOS/occupation-specific technical tasks at the appropriate skill level. Stays informed of the latest developments in field.

16. **Common Task Knowledge and Skill.** Possesses the necessary knowledge and skill to perform common tasks at the appropriate skill level (e.g., land navigation, field survival techniques, and nuclear, biological, and chemical [NBC] protection).

17. **Safety Consciousness.** Follows safety guidelines and instructions. Checks the behavior of others to ensure compliance.

18. **Computer Skills.** Understands computer systems, operating systems (e.g., Unix, Windows NT, and Army specific systems) and applications. Can perform routine troubleshooting of computer systems and applications.

19. **Motivating, Leading, and Supporting Individual Subordinates.** Recognizes, encourages, and rewards effective performance of individual subordinates. Corrects unacceptable conduct. Communicates reasons for actions and listens effectively to subordinates one-on-one. Fosters loyalty and commitment.

20. **Directing, Monitoring, and Supervising Individual Subordinates.** Works with subordinates one-on-one to assign tasks and set individual goals for work and assignments. Ensures that assignments are clearly understood. Monitors individual subordinate performance and gives appropriate feedback.

21. **Training Others.** Evaluates and identifies individual or unit training needs. Institutes formal or informal programs to address training needs. Develops others by providing appropriate work experiences. Guides and tutors subordinates on technical matters.

22. **Relating to and Supporting Peers.** Treats peers in a courteous, respectful, and tactful manner. Provides help and assistance to others. Backs up and fills in for others when needed. Works effectively as a team member.

23. **Team Leadership.** Communicates team goals and organizes and rewards effective teamwork. Leads the team to adapt quickly when missions change and keeps team focused on new goals. Resolves conflicts among team members. Shares relevant information with team members.

24. **Concern for Soldier Quality of Life.** Is aware of subordinates' off-duty needs and constraints. Is sensitive to others' priorities, interests, and values, and tries to assist subordinates in making their personal and family life better.

25. **Cultural Tolerance.** Demonstrates tolerance and understanding of individuals from other cultural and social backgrounds, both in the context of the diversity of U.S. Army personnel and interactions with foreign nationals during deployments or when training for deployment.

Table 1.3. NCO21 Knowledges, Skills, and Aptitudes (KSAs) and Performance Requirements (Continued)

26.	**Modeling Effective Performance.** Acts in ways that consistently serve as a model for what effective performance should be like, be it technical performance, military bearing, commitment to the Army, support for the Army mission, or performance under stressful or adverse conditions. Can consistently set an example for others to follow.
27.	**Level of Effort and Initiative on the Job.** Demonstrates high effort in completing work. Takes independent action when necessary. Seeks out and willingly accepts responsibility and additional challenging assignments. Persists in carrying out difficult assignments and responsibilities.
28.	**Adherence to Regulations, Policies, and Procedures.** Adheres to policies and follows prescribed procedures in carrying out duties and assignments.
29.	**Level of Integrity and Discipline on the Job.** Maintains high ethical standards. Does not succumb to peer pressure to commit prohibited, harmful, or questionable acts. Demonstrates trustworthiness and exercises effective self-control. Understands and accepts the basic values of the Army and acts accordingly.
30.	**Adaptability.** Can modify behavior or plans as necessary to reach goals or to adapt to changing goals. Is able to maintain effectiveness when environments, tasks, responsibilities, or personnel change. Easily commits to learning new things when the technology, mission, or situation requires it.
31.	**Physical Fitness.** Meets Army standards for weight, physical fitness, and strength. Maintains health and fitness to meet deployability and field requirements as well as the physical demands of the daily job.
32.	**Military Presence.** Presents a positive and professional image of self and the Army even when off duty. Maintains proper military appearance.
*33.	**Information Management.** Effectively monitors, interprets, and redistributes digital display information (as well as printed and orally delivered information) from multiple sources to multiple recipients. Sorts, classifies, combines, excludes, and presents information so that it is useable by others. Does not readily succumb to information overload.
*34.	**Selfless Service Orientation.** Commits to the greater good of the team or group. Puts organizational goals ahead of individual goals as required.
*35.	**General Self-Management Skill.** Uses appropriate strategies to self-manage the full range of own work and non-work responsibilities (e.g., work assignments, personal finances, family). Such strategies include setting both long- and short-term goals, allocation of effort and personal resources to goal priorities, and assessing one's own performance. Works effectively without direct supervision, but seeks help and advice from others when appropriate.
*36.	**Self-Directed Learning Skill.** Has a clear goal of maintaining continuous learning and training over entire career. Is proficient at determining personal training needs, planning education and training experiences to meet them, and evaluating own training success. Uses efficient personal learning strategies (e.g., organizing the material to be learned, and practicing the new skills in an appropriate context).
*37.	**Knowledge of the Inter-Relatedness of Units.** Is capable of analyzing how goals and operations of own unit are inter-related with other units and systems, and how one unit's actions affect the performance of other units. Can see the larger strategic picture and interpret how one's own unit relates to it.
*38.	**Management and Coordination of Multiple Battlefield Functions.** Can individually apply and effectively integrate and coordinate multiple battlefield functions such as direct and indirect fires, communications, intelligence, and combat service support to achieve tactical goals.

Note. KSAs/performance requirements that are particularly relevant to one or both future eras, but not necessarily for the baseline era, are noted with an asterisk.

Identification of Alternative Measurement Methods

Literature Review

A literature review was conducted to identify existing instruments that might be used to measure each KSA (predictor) and performance requirement (criterion) identified in Phase II of the NCO21 project. The relevant literature comprised research studies, instrument development projects, and test publishers that have developed or used measures potentially applicable to the NCO21 KSAs and performance requirements. Consequently, information from relevant Army and other Department of Defense (DoD) research and practice, private-sector research and development, and test-publisher products was surveyed. Particular attention (especially regarding job performance measurement) was paid to research and practice in the military services.

Two sets of tables were constructed to help summarize the information collected in the literature review. The first table contained the potential measures for each of the 38 KSAs and the second table contained comparable information for the performance requirements. Two tables were required because, although the performance requirements are subsumed within the KSA list, there was some difference in potential measurement methods. Specifically, the criterion measures are for research use only (so, for example, a complicated work sample or simulation could be considered), whereas the predictor measures must have the potential to be administered operationally. The tables identified "best bet" measures for each KSA/performance requirement. Best bet measures were identified based on (a) the degree to which the measure was developed for the Army or other military service (i.e., the more Army-specific the better), (b) the judged conceptual and psychometric quality of the measure, and (c) the judged "feasibility" of using the measure in an NCO selection system (this latter judgment pertained to the KSA measures only). It was also noted that, other things being equal, the Army would likely prefer performance-based over trait-based KSA measures. For example, the preference would be to use a job performance-based assessment of conscientious work behavior rather than a personality inventory assessment of conscientiousness. The overall objective was to portray a reasonable picture of available potential measures that neither narrowed the possibilities too much nor included too many irrelevant and distracting possibilities. A package of 1- to 2-page summaries of over 50 potentially relevant measures referred to in the tables was prepared. These summaries drew heavily on prior work sponsored by ARI (Russell et al., 1995).

Expert Psychologists Panel

The results of the project team's literature review work were presented to a working group of psychologists experienced in predictor and criterion measurement. In addition to scientists from HumRRO and ARI, the group included three expert consultants who were selected based on their exceptional technical qualifications, experience, and breadth of expertise. The expert consultants were tasked with (a) reviewing in detail the project team's progress on the literature review, (b) adding or deleting measures and measurement methods from the lists of possibilities prepared by project staff, and (c) providing general counsel to the project team about strategies for achieving the desired measurement goals. They reviewed a package containing the summary tables and instrument descriptions developed by project staff. The consultants then participated in a 1-day meeting with HumRRO and ARI staff to discuss measurement of each KSA and performance criterion in turn.

Selection of Measurement Methods

Up to this point, the strategy was to consider all reasonable (broadly defined) measurement methods applicable to one or more of the NCO21 KSAs or performance requirements. After the expert psychologists' input was received, however, it became necessary to determine which measures or measurement methods would actually be used in the NCO21 validation data collection. The "best bet" possibilities were reviewed to determine the set of measures that most closely met the following criteria:

- Coverage of highest priority KSAs as determined by the Phase II expert panels.
- Coverage of performance requirements.
- Anticipated reliability and validity in an operational context.
- Reasonable development and validation costs.
- Suitability of KSA measures for a large-scale operational promotion system.

Project staff developed a set of recommended measures and prepared a brief discussion paper for each measure. Each discussion paper provided preliminary thinking about what the recommended measure would look like (if it had to be developed), what KSAs or performance requirements it might cover, what issues would need to be addressed if it was adopted, and how it would be used. A decision meeting involving HumRRO and ARI staff was held to review recommendations and reach final decisions about what instruments would be pursued. The measures listed in Table 1.4 were selected using the above criteria.

Table 1.4. NCO21 Criterion and Predictor Measures

Criterion Measures
Supervisor ratings (Observed Performance Rating Scales and Expected Future Performance Rating Scales)
Aptima computerized simulation
Predictor Measures
Situational Judgment Test (SJT)
Self-report archival information (collected on the Personnel File Form-21)
Self-report accomplishments (collected on the Experience and Activities Record)
Structured interview (administered by trained senior NCOs)
Armed Services Vocational Aptitude Battery (ASVAB)
Assessment of Individual Motivation (AIM)
Biographical Information Questionnaire (BIQ)

Note that Table 1.4 contains an entry entitled *Aptima computerized simulation.* This computerized simulation exercise is being developed under a companion project sponsored by ARI and conducted by Aptima Human-Centered Engineering, Inc. HumRRO staff are working with Aptima researchers to help ensure that the simulation exercise elicits criterion constructs of interest to the NCO21 research program. Because the simulation is in the early stages of development, it will not be available to administer to most of the NCO21 validation data

collection participants. To the extent that it can be administered to a subset of the sample, however, it will provide another important source of performance criterion information.

Overview of Measure Development

Three of the selected measures (ASVAB, AIM, and BIQ) required little or no development. These are tests already used operationally in other contexts, so at most, additional experimental items were added for the NCO21 research. For the other measures, every effort was made to build upon prior work whenever possible. Of particular value was material and experience gained from Project A (J. Campbell & Knapp, 2001) and the Expanding the Concept of Quality in Personnel (ECQUIP) project. Project A was conducted in the 1980s and focused on the development and validation of tools to select and classify applicants into entry-level Army enlisted jobs. ECQUIP was conducted in the 1990s and focused on predictors of Army NCO (grades E5 through E8) performance. As an example of how this earlier work contributed to the present research, rating instruments and rater training procedures developed and used in these earlier projects were used as a starting point for the NCO21 Observed Performance Rating Scales instrument and rater training procedures.

The steps followed to prepare each measure varied considerably across the measures and will be described in more detail in subsequent chapters. Here we provide a brief overview of the data collection efforts that supported instrument development and field-testing activities.

Sites Supporting Instrument Development and Pilot Testing

The research team worked within the constraints of pre-determined FY 2000 research support requests to obtain the input required for instrument development and field-testing. Army sites scheduled to provide troop support early in the year were used for instrument development and pilot testing. These sites included Forts Campbell, Bragg, Riley, and Knox, as well as the U.S. Sergeants Major Academy (USASMA). Senior NCOs (E7s and above) were provided at Fort Knox and USASMA. The remaining sites provided the E4 through E9 level personnel to support instrument development efforts. Generally, project staff developed interactive workshop activities for the E7 through E9 NCOs (e.g., to generate Situational Judgment Test items) and administered draft instruments to E4, E5, and E6 level participants in classroom settings.

Field Test Data Collections

Three Army installations scheduled to provide troop support somewhat later in the year (Forts Carson, Leonard Wood, and Stewart) provided field test data. The purpose of the field test was to evaluate and finalize the new measures and to try-out and refine data collection procedures.

Data collection staff participated in intensive training that included an introduction to the project and careful review of the data collection procedures detailed in the data collection staff manual that had been prepared. Staff involved in the collection of supervisor ratings and/or the administration of the structured interviews were given additional training on the more complex procedures associated with those measures. For example, they practiced how to administer supervisor rater training and worked on strategies to maximize the amount and quality of data collected. Several staff members were also tasked with providing training for the senior NCOs who would be administering the structured interviews to participating soldiers.

Table 1.5 shows the measures that were administered in the field test, by grade. E4 level participants took all of the predictor measures, but were assessed by none of the criterion measures (performance ratings) because these measures were developed to assess E5 and E6 level performance. E5 soldiers took all of the predictor measures and were assessed by both criterion measures. E6 soldiers took most of the predictor measures and were also assessed by both criterion measures. Although the predictor measures can be administered to E6 soldiers, they are targeted to the E4 and E5 levels. During the test sessions, some E4 and E5 soldiers were selected to participate in the semi-structured interview. The supervisor ratings were collected in concurrent sessions.

Table 1.5. Field Test Instruments

Instrument	E4	E5	E6
Background Information Form	✓	✓	✓
Situational Judgment Test (SJT)	✓	✓	✓
Situational Judgment Test – X (SJT-X)			✓
Experiences and Activities Record (ExAct)	✓	✓	✓
Personnel File Form-21 (PFF21)	✓	✓	✓
Assessment of Individual Motivation (AIM)	✓	✓	
Biographical Information Questionnaire (BIQ)	✓	✓	
Semi-Structured Interview	✓	✓	
Observed Performance Rating Scales		✓	✓
Expected Future Performance Rating Scales		✓	✓

Note. The Background Information Form was for administrative purposes only, eliciting social security numbers and other identifying information. The SJT-X, a variation of the SJT, was administered at only one test site. The rating scales were completed by the soldiers' supervisors.

The number of soldiers who could be interviewed depended on the number of interviewer teams provided by the installation. Supervisor ratings were collected for the E5 and E6 soldiers participating in the data collection. All research participants completed a Background Information Form eliciting descriptive information (e.g., grade, test location).

Field Test Database

Data were collected from 513 soldiers: 189 from Fort Carson (37%), 101 from Fort Leonard Wood (20%), and 223 from Fort Stewart (43%). Table 1.6 shows sample sizes by grade, sex, race/ethnic group, and type of military occupational specialty (MOS). When sample sizes permitted, score differences across subgroups were examined. Sample sizes for the interviews and the supervisor ratings are provided in Chapters 2 and 5, respectively.

Overview of Report

Chapters 2-6 present in detail the development and field-testing of the NCO21 criterion and predictor measures. Although sample sizes with complete data on all instruments were insufficient to conduct the types of criterion-related validity analyses planned for the validation effort, Chapter 7 provides a preliminary assessment of the inter-relationships among the different

predictor measures and correlations between the predictors and the performance criterion ratings. Finally, Chapter 8 summarizes the final recommendations for predictor and criterion measurement in the NCO21 validation data collection.

Table 1.6. Field Test Subgroup Sample Sizes

Grade	Sex		Race			MOS		
	Male	Female	White	Black	Other	C	CS	CSS
E4	162	35	120	45	32	50	45	99
E5	175	29	124	52	28	46	56	99
E6	84	9	50	34	9	30	26	38
Total	421	73	294	131	69	126	127	236

Note. C = Combat, CS = Combat Service, CSS = Combat Service Support. Cases with missing data are excluded from these counts.

CHAPTER 2: SUPERVISOR RATINGS

Background

As discussed in Chapter 1, supervisor ratings of E5 and E6 level performance were selected to serve as the primary criterion measurement method for the NCO21 validation effort. Another type of criterion measure, the computerized work sample simulation being developed under contract to Aptima Human-Centered Engineering, Inc., would be available to administer to a small subset of E5 and E6 soldiers participating in the NCO21 validation.

Two rating instruments were designed–one to assess observed job performance and another to predict how well soldiers would perform under expected future Army conditions. Specifically, the Observed Performance Rating Scales were used to collect supervisor ratings of subordinate E5 and E6 soldiers' "typical" job performance in a broad spectrum of performance areas (i.e., all 27 NCO21 performance requirements listed in Table 1.3), overall soldier performance, and potential performance as a senior NCO (i.e., E7, E8, or E9). The Expected Future Performance Rating Scales were used to obtain supervisor ratings on how well E5 and E6 soldiers could be expected to perform in several scenarios describing conditions predicted to occur in the future Army. These measures are based on the Project A model that views job performance as a multi-dimensional construct in which multiple attributes, outcomes, or factors are indicative of job performance (C. Campbell et al., 1990). The goal of the supervisor rating instruments is to describe and evaluate soldiers on the performance requirements that constitute effective performance across all Army jobs. Such performance requirements have been termed "Army-wide" criterion factors in previous research (Borman, Motowidlo, Rose, & Hanser, 1985). The development of each of the supervisor rating instruments is described independently below.

Observed Performance Rating Scales

Instrument Development Process

The Observed Performance Rating Scales were modeled after the Project A second-tour Army-wide performance rating scales (J. Campbell & Knapp, 2001). Each performance dimension is assessed on a 7-point scale, and each rating scale has three definition-based anchor statements that describe low, moderate, and high performance. A total of 16 scales on the first version of the rating instrument were derived from existing measures–10 from the Project A scales (which were administered to E5 NCOs) and 6 from the ECQUIP project (Peterson et al., 1997). The ECQUIP scales were designed for E5-E8 NCOs. The anchor statements in the ECQUIP scales were modified to reflect the format of the Project A scales. In most cases, relatively few substantive changes were required to make the Project A and ECQUIP scales suitable for the NCO21 instrument. Some performance requirements were not covered by existing scales, however, so project staff drafted new scales for these 11 performance requirements: (a) General Self-Management Skill, (b) Common Task Knowledge and Skill, (c) Information Management, (d) Computer Skills, (e) Cultural Tolerance, (f) Modeling Effective Performance, (g) Management/Coordination of Multiple Battlefield Functions, (h) Knowledge of the Inter-Relatedness of Units, (i) Team Leadership, (j) Problem-Solving/Decision Making, and (k) Selfless Service Orientation.

In preparation for the first pilot test administration, a test administrator's script and instructions for the supervisor raters were drafted. Rater training included an overview of the instrument and performance areas to be rated, directions to supervisors on how to mark their ratings, and examples of common rating errors to avoid (e.g., leniency, halo, recency).

The project team reviewed and revised the ratings instrument and supporting materials several times before the first pilot test administration. For example, some anchor statements were made more concise, as the amount of reading required was of concern. Also, two additional scales were added to the rating instrument between the first and second pilot test: (a) a rating of overall soldier effectiveness, and (b) a rating of potential effectiveness as a senior NCO. As with the performance requirement scales, these scales contained 7-point response options, with anchor levels describing low, moderate, and high performance. These scales were taken directly from the Project A second-tour Army-wide performance ratings instrument. Thus, the final pilot test instrument consisted of 29 rating scales – one for each of the 27 performance requirements identified as relevant for 21st-century NCOs (see Table 1.3), one overall performance scale, and one senior NCO potential scale.

Pilot Testing the Prototype

The prototype Observed Performance Rating Scales and associated rater training procedures were tried out and revised in an iterative fashion across three pilot tests. An abbreviated version of the scales was also administered to a small sample of raters to explore concerns about the basis for some of the ratings.

In the three primary pilot tests, E6-E9 NCOs (n = 137) were instructed to rate the performance of two unspecified subordinate soldiers at the E5 or E6 level. They listened to the rater training provided by a test administrator and read the instructions in their rating booklet. They were also asked to provide feedback about the instrument and the rater training.

The first two pilot tests showed little evidence of leniency or central tendency in the NCO ratings. All item means were between 4.2 and 5.3 on the 7-point scale. There was more evidence of leniency in the third pilot test, where the item means ranged from 4.8 to 5.8. The results from all the pilot test sites suggested some halo, and ratings were rendered for some areas that probably had not been observed by all of the raters (e.g., Information Management, Management/Coordination of Multiple Battlefield Functions). Although raters were somewhat less likely to rate these areas than others, the number of ratings provided was still higher than anticipated given that most soldiers in today's Army are not required to perform in these areas.

To examine this issue further, an abbreviated version of the rating scales was pre-tested on a convenience sample of three company commanders (O3) and three E7 NCOs. These raters had experience in digitized jobs and thus should have been more likely to be able to rate subordinates on futuristic performance requirements. The abbreviated instrument included rating scales for the more future-oriented requirements (e.g., Computer Skills, Information Management, Knowledge of the Inter-relatedness of Units, and Management/Coordination of Multiple Battlefield Functions). We added two questions to the abbreviated instrument to ascertain the basis for the ratings. One question asked whether each rating was based on (a) actual job performance or (b) how raters thought the soldier *might* perform (without actually having observed performance). The answers

indicated that even these relatively future-oriented supervisors generalized their ratings for areas they observed to areas they did not observe. Raters were also asked to describe the basis for their ratings of each area that was not directly observed. The results suggested that most of the inferred ratings were based on factors such as transferable skills demonstrated in other aspects of performance (e.g., "his ability to understand and operate in separate battlefield functions would lead me to believe that he has the ability to coordinate them when the job asked him to do so"). In an effort to reduce the extent to which raters would infer their supposedly "observed" performance ratings, the rater training script was changed to encourage raters to use the "cannot rate" option as needed.

Preparation for the Field Test

As indicated previously, the Observed Performance Rating Scales and associated rater training were revised in an iterative fashion. Changes were made to all elements of the process – the rating scales anchors, written instructions, and oral test administrator script. Moreover, the pilot tests involved rating two unspecified E5 and/or E6 soldiers so the instrument and instructions needed to be changed for the field test to accommodate ratings for specific soldiers. Using the Project A rating scales as a model, the format of the Observed Performance Rating Scales and instructions were changed to allow for ratings of up to five soldiers per rating booklet. The instrument was also converted to a machine-scoreable format.

Field Test Administration

The Observed Performance Rating Scales were field tested at three Army installations. Two supervisors were requested for each E5/E6 soldier participating in the field test. Supervisors could include a soldier's official first line supervisor, second line supervisor, and/or any outranking NCO with whom the soldier routinely works. Raters must have worked with the soldier for at least one month. The goal was to obtain ratings from two supervisors per soldier participating in the predictor measure test sessions.

Project staff provided training on how to use the instrument. While the supervisors were completing their ratings, the test administrators watched for supervisors who were clearly not reading the anchor definitions when making their ratings. Such cases were pointed out in a general manner to the entire group, and the raters were reminded to use the rating scales when making the ratings.

Field Test Results

The field test data for the Observed Performance Rating Scales included scores on the 27 performance requirements, an overall performance score, and a senior NCO potential score, for a total of 29 rating scale scores. Ratings on each item ranged from 1 (low) to 7 (high).

Descriptive Statistics

A total of 211 soldiers (n_{E5} = 137, n_{E6} = 74) were rated by at least one supervisor; 72 (34%) of these soldiers were rated by two or more supervisors. The analyses excluded data from supervisors who had worked with the soldier for less than one month (based on information obtained from the Supervisor Background Information Form).

Table 2.1 shows the mean ratings for each individual rating scale. Although the mean values (4.4-5.7) suggest some leniency in the ratings, the amount of variability in the ratings suggests supervisors were able to discriminate among soldiers on each scale. The data also show a reasonable response pattern, such that the performance requirements expected to be observed by the fewest number of supervisors received the greatest amount of "cannot rate" responses (e.g., Information Management, Management/Coordination of Multiple Battlefield Functions).

Table 2.1. Descriptive Statistics for the Observed Performance Rating Scales

Scale	*M*	*SD*	*n*
1. MOS/Occupation-Specific Knowledge and Skill	5.37	1.15	204
2. Common Task Knowledge and Skill	5.37	1.19	203
3. Computer Skills	4.70	1.48	181
4. Safety Consciousness	5.69	1.00	206
5. Writing Skill	4.71	1.20	189
6. Oral Communication Skill	5.18	1.22	210
7. Level of Effort and Initiative on the Job	5.22	1.39	211
8. Adaptability	5.09	1.18	203
9. General Self-Management Skill	5.11	1.39	207
10. Self-Directed Learning Skill	4.79	1.37	204
11. Adherence to Regulations, Policies, and Procedures	5.43	1.27	211
12. Level of Integrity and Discipline on the Job	5.50	1.26	210
13. Physical Fitness	5.07	1.50	209
14. Military Presence	5.14	1.22	210
15. Relating to and Supporting Peers	5.05	1.19	211
16. Team Leadership	5.08	1.23	197
17. Cultural Tolerance	5.54	1.10	187
18. Selfless Service Orientation	5.06	1.21	210
19. Motivating, Leading, and Supporting Individual Subordinates	4.73	1.32	201
20. Directing, Monitoring, and Supervising Individual Subordinates	4.80	1.28	190
21. Modeling Effective Performance	4.88	1.18	205
22. Concern for Soldier Quality of Life	5.23	1.15	195
23. Training Others	4.90	1.24	195
24. Knowledge of the Inter-Relatedness of Units	4.89	1.28	198
25. Management/Coordination of Multiple Battlefield Functions	4.43	1.33	147
26. Problem-Solving/Decision Making Skill	5.05	1.25	205
27. Information Management	4.89	1.22	188
Composite	5.08	0.84	211
Overall Effectiveness	5.07	1.15	206
Senior NCO Potential	4.90	1.36	201

Note. The *n* varies because of use of the "cannot rate" option.

A composite observed performance rating was computed from the mean of the 27 performance requirement ratings. Table 2.2 shows that this composite score was highly correlated with the single item rating for overall effectiveness, $r = .85$, $p < .01$. The table also shows the correlations of the scores on each rating scale with the composite, overall effectiveness, and senior NCO potential scores. Recall that raters were consistent within themselves in their ratings of the 27 performance areas (i.e., halo); this likely contributed to the high correlations among the Observed Performance Rating Scale scores.

Table 2.2. Correlations Between the Individual Performance Requirement Rating Scales and Global Scores

Scale	Composite	Overall Effectiveness	Senior NCO Potential
1. MOS/Occupation-Specific Knowledge and Skill	.75**	.70**	.64**
2. Common Task Knowledge and Skill	.72**	.60**	.57**
3. Computer Skills	.29**	.16*	.13
4. Safety Consciousness	.62**	.41**	.37**
5. Writing Skill	.57**	.49**	.42**
6. Oral Communication Skill	.58**	.45**	.47**
7. Level of Effort and Initiative on the Job	.75**	.67**	.63**
8. Adaptability	.71**	.58**	.50**
9. General Self-Management Skill	.72**	.64**	.63**
10. Self-Directed Learning	.62**	.45**	.48**
11. Adherence to Regulations, Policies, and Procedures	.65**	.52**	.49**
12. Level of Integrity and Discipline on the Job	.75**	.62**	.55**
13. Physical Fitness	.56**	.51**	.50**
14. Military Presence	.65**	.54**	.52**
15. Relating to and Supporting Peers	.67**	.55**	.48**
16. Team Leadership	.79**	.69**	.64**
17. Cultural Tolerance	.35**	.29**	.26**
18. Selfless Service Orientation	.68**	.65**	.55**
19. Motivating, Leading, and Supporting Individual Subordinates	.80**	.68**	.70**
20. Directing, Monitoring, and Supervising Individual Subordinates	.77**	.68**	.61**
21. Modeling Effective Performance	.82**	.79**	.72**
22. Concern for Soldier Quality of Life	.64**	.47**	.44**
23. Training Others	.69**	.61**	.55**
24. Knowledge of the Inter-Relatedness of Units	.67**	.56**	.45**
25. Management/Coordination of Multiple Battlefield Functions	.70**	.64**	.60**
26. Problem-Solving/Decision Making Skill	.76**	.68**	.69**
27. Information Management	.68**	.54**	.54**
Composite	1.00	--	--
Overall Effectiveness	.85**	1.00	--
Senior NCO Potential	.79**	.82**	1.00

*$p < .05$. **$p < .01$.

Subgroup Differences

Table 2.3 shows subgroup differences in soldier ratings by gender, race, grade, and MOS type (i.e., combat, combat support, and combat service support). The analyses revealed no significant differences in composite observed performance scores based on gender, race, or MOS type. As expected, there was a significant difference in the composite score by grade, such that E6 soldiers were rated higher than E5 soldiers, $p < .001$.

Table 2.3. Subgroup Differences for the Composite Observed Performance Rating Scale Score

Group	*n*	*M*	*SD*	Effect Size	*p*
Gender					
Female	22	5.15	0.85	0.10	.656
Male	182	5.07	0.81		
Race					
Black	58	5.06	0.82	-0.02	.894
White	123	5.08	0.83		
Pay Grade					
E6	74	5.45	0.72	0.69	<.001
E5	137	4.88	0.83		
MOS Type					
Combat Support	62	5.10	0.69	-0.17	.349
Combat	51	5.23	0.77		
Combat Service Support	88	4.95	0.98	-0.36	.070
Combat	51	5.23	0.77		
Combat Service Support	88	4.95	0.98	-0.21	.291
Combat Support	62	5.10	0.69		

Note. Subgroup differences in composite score. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* referent group. Referent groups (e.g., Whites) are listed second in each pair. Care should be exercised when interpreting the statistics because unequal cell sizes and interaction effects were not taken into account.

Dimensionality

The 27 performance requirement ratings were correlated to assess their inter-relationships (see Table 2.4). Correlations ranged from -.03 to .72 and most were significant at the $p < .01$ level. The Computer Skills ratings were correlated the lowest with the other performance requirements (r = -.03 to .35).

Confirmatory factor analysis and exploratory factor analysis were performed independently to determine if the item (i.e., individual performance requirement rating scale) scores could be consolidated into dimensions. Confirmatory factor analyses were conducted to test two a priori models: the Project A 6-factor performance model (J. Campbell & Knapp, 2001) and the NCO21 performance component model (i.e., a rational grouping of the 27 performance requirements into 14 performance "components"). Neither model yielded meaningful results. The first did not fit well and the second was too complex (i.e., too many parameters relative to the sample size).

Table 2.4. Inter-Item Correlations Among Observed Performance Ratings

	1	2	3	4	5	6	7	8	9	10
1. MOS/Occupation-Specific Knowledge and Skill										
2. Common Task Knowledge and Skill	.59**									
3. Computer Skills	.07	-.01								
4. Safety Consciousness	.32*	.33**	.13*							
5. Writing Skill	.42**	.34**	.35**	.37**						
6. Oral Communication Skill	.40**	.50**	.25**	.27**	.44**					
7. Level of Effort and Initiative on the Job	.58**	.52**	.12*	.39**	.32**	.38**				
8. Adaptability	.50**	.53**	.08	.41**	.38**	.47**	.57**			
9. General Self-Management Skill	.52**	.49**	.15*	.39**	.36**	.50**	.58**	.53**		
10. Self-Directed Learning Skill	.47**	.46**	.20**	.43**	.30**	.28**	.46**	.35**	.47**	
11. Adherence to Regulations, Policies, and Procedures	.49**	.43**	.03	.53**	.25**	.29**	.51**	.44**	.45**	.40**
12. Level of Integrity and Discipline on the Job	.49**	.43**	.18**	.53**	.32**	.36**	.59**	.57**	.56**	.39**
13. Physical Fitness	.44**	.38**	.05	.41**	.32**	.30**	.35**	.36**	.34**	.33**
14. Military Presence	.46**	.39**	.12	.39**	.38**	.32**	.42**	.43**	.43**	.48**
15. Relating to and Supporting Peers	.47**	.37**	.16*	.36**	.35**	.27**	.47**	.49**	.39**	.30**
16. Team Leadership	.55**	.56**	.13	.50**	.41**	.41**	.59**	.52**	.57**	.38**
17. Cultural Tolerance	.17*	.18**	-.03	.26**	.19**	.13*	.26**	.30**	.17**	.11
18. Selfless Service Orientation	.55**	.45**	.01	.39**	.23**	.28**	.61**	.54**	.50**	.39**
19. Motivating, Leading, and Supporting Individual Subordinates	.54**	.56**	.20**	.49**	.34**	.43**	.63**	.50**	.60**	.54**
20. Directing, Monitoring, and Supervising Individual Subordinates	.60**	.62**	.24**	.43**	.39**	.41**	.58**	.44**	.55**	.47**
21. Modeling Effective Performance	.62**	.60**	.23**	.49**	.42**	.47**	.62**	.56**	.60**	.46**
22. Concern for Soldier Quality of Life	.46**	.48**	.18**	.48**	.42**	.28**	.39**	.41**	.38**	.39**
23. Training Others	.58**	.51**	.12	.37**	.34**	.40**	.41**	.38**	.46**	.39**
24. Knowledge of the Inter-Relatedness of Units	.49**	.46**	.28**	.30**	.26**	.36**	.48**	.54**	.42**	.38**
25. Management/Coordination of Multiple Battlefield Functions	.57**	.59**	.25**	.29**	.42**	.50**	.41**	.40**	.42**	.35**
26. Problem-Solving/Decision Making Skill	.59**	.53**	.26**	.44**	.40**	.41**	.63**	.53**	.60**	.37**
27. Information Management	.53**	.47**	.31**	.35**	.51**	.35**	.52**	.40**	.39**	.39**

Table 2.4. Inter-Item Correlations Among Observed Performance Ratings (Continued)

	11	12	13	14	15	16	17	18	19	20
1. MOS/Occupation-Specific Knowledge and Skill										
2. Common Task Knowledge and Skill										
3. Computer Skills										
4. Safety Consciousness										
5. Writing Skill										
6. Oral Communication Skill										
7. Level of Effort and Initiative on the Job										
8. Adaptability										
9. General Self-Management Skill										
10. Self-Directed Learning Skill										
11. Adherence to Regulations, Policies, and Procedures										
12. Level of Integrity and Discipline on the Job	.65**									
13. Physical Fitness	.29**	.40**								
14. Military Presence	.38**	.43**	.50**							
15. Relating to and Supporting Peers	.43**	.64**	.32**	.38**						
16. Team Leadership	.56**	.61**	.35**	.48**	.59**					
17. Cultural Tolerance	.20**	.24**	.15*	.24**	.35**	.36**				
18. Selfless Service Orientation	.47**	.54**	.32**	.44**	.51**	.48**	.32**			
19. Motivating, Leading, and Supporting Individual Subordinates	.52**	.53**	.40**	.53**	.48**	.72**	.22**	.56**		
20. Directing, Monitoring, and Supervising Individual Subordinates	.45**	.50**	.40**	.45**	.40**	.64**	.16*	.48**	.71**	
21. Modeling Effective Performance	.57**	.62**	.48**	.53**	.54**	.67**	.19**	.58**	.70**	.63**
22. Concern for Soldier Quality of Life	.33**	.43**	.32**	.41**	.44**	.51**	.25**	.38**	.55**	.43**
23. Training Others	.41**	.41**	.33**	.43**	.42**	.57**	.17*	.44**	.61**	.58**
24. Knowledge of the Inter-Relatedness of Units	.34**	.47**	.29**	.41**	.50**	.45**	.30**	.45**	.42**	.44**
25. Management/Coordination of Multiple Battlefield Functions	.16*	.42**	.43**	.47**	.40**	.46**	.16*	.41**	.49**	.57**
26. Problem-Solving/Decision Making Skill	.47**	.58**	.34**	.31**	.51**	.65**	.18**	.51**	.62**	.60**
27. Information Management	.28**	.35**	.28**	.36**	.37**	.45**	.18**	.39**	.50**	.50**

Table 2.4. Inter-Item Correlations Among Observed Performance Ratings (Continued)

	21	22	23	24	25	26
1. MOS/Occupation-Specific Knowledge and Skill						
2. Common Task Knowledge and Skill						
3. Computer Skills						
4. Safety Consciousness						
5. Writing Skill						
6. Oral Communication Skill						
7. Level of Effort and Initiative on the Job						
8. Adaptability						
9. General Self-Management Skill						
10. Self-Directed Learning Skill						
11. Adherence to Regulations, Policies, and Procedures						
12. Level of Integrity and Discipline on the Job						
13. Physical Fitness						
14. Military Presence						
15. Relating to and Supporting Peers						
16. Team Leadership						
17. Cultural Tolerance						
18. Selfless Service Orientation						
19. Motivating, Leading, and Supporting Individual Subordinates						
20. Directing, Monitoring, and Supervising Individual Subordinates						
21. Modeling Effective Performance						
22. Concern for Soldier Quality of Life	.46**					
23. Training Others	.54**	.47**				
24. Knowledge of the Inter-Relatedness of Units	.50**	.34**	.46**			
25. Management/Coordination of Multiple Battlefield Functions	.54**	.50**	.47**	.46**		
26. Problem-Solving/Decision Making Skill	.58**	.41**	.49**	.49**	.55**	
27. Information Management	.48**	.48**	.46**	.50**	.58**	.61**

Note. $n = 134 - 211$.
$^{*}p < .05$. $^{**}p < .01$

The exploratory factor analyses were conducted using the maximum likelihood extraction method with oblique rotation. Models containing between two and eight factors were tested. The favored model yielded five factors. Table 2.5 shows this model, modified to form seven dimensions. For theoretical reasons, one of the original factors was separated into two factors (now referred to as Dimensions 1 and 2). Oral Communication was isolated as a separate single-item dimension because its loadings were unclear and because it conceptually did not belong with the factor on which it had the highest loading. In addition, three scale scores were not included in the model because the score either loaded highly on more than one dimension (Modeling Effective Performance), did not load on any factor (General Self-Management Skill), or was considered sample-specific (MOS/Occupation-Specific Knowledge and Skill). Table 2.5 also shows the correlations between the seven dimension scores and the composite score. All dimension scores were correlated highly with the composite score, $p < .001$.

Reliability Estimates

The internal consistency reliability (Cronbach's alpha) was computed for the composite score and the dimension scores. For the composite score, the alpha was .94; the internal consistency reliability estimates for the dimensions are presented in Table 2.5. The reliability estimates for the six dimensions (Dimension 7 was a single item) ranged from .64 (Getting Along with Others) to .87 (Leadership). These results suggest that, in general, the most parsimonious summary score is the composite score; however, the dimensions work reasonably well if further differentiation is desired.

Interrater reliability was estimated using soldiers who had at least two supervisor raters. For each soldier, one supervisor was assigned to the first rater group and another was assigned to the second rater group. Reliabilities were estimated by computing the (a) correlation between the two groups of ratings for each scale, composite, and dimension (r) and (b) intraclass correlation coefficients (*ICCs*; Shrout & Fleiss, 1979, *ICC*[3,1] assessing consistency across a fixed set of raters). The two estimates yielded almost identical results, as shown in Table 2.6. Interrater reliability was fairly high for several areas (e.g., Physical Fitness, MOS/Occupation-Specific Knowledge and Skill, Military Presence) but was low (i.e., near zero) for other areas (e.g., Writing Skill, Adaptability, General Self-Management Skill). The interrater reliability was relatively good for the composite score and for several of the dimension scores (e.g., Dimensions 1, 2, and 4).

Preparation for the Validation Data Collection

The composite score yields high internal consistency reliability and it was used as the primary score for the field test analyses. Because the interrater reliability estimates are lower than desired, several steps were taken to enhance the utility of the tool in preparation for the validation data collection. The most potentially effective changes pertain to (a) the number and quality of the raters per ratee and (b) a reduction in the number of required ratings. With regard to the number and quality of the raters, the troop support requests for the validation data collection more clearly specify the need for two supervisors per soldier. Pre-coordination efforts with test site personnel will also focus much more heavily on the need for supervisors who have had sufficient experience observing their soldier's performance (i.e., at least 3 months working with the soldier).

Table 2.5. Seven Dimensions Based on 5-Factor Exploratory Factor Analysis

Number	Dimension/Item	Internal Consistency Reliability	Correlation with Composite
Dimension 1	**Fosters Improvement of Self and Others**	.73	.84
Scale 10	Self-Directed Learning Skill		
Scale 13	Physical Fitness		
Scale 14	Military Presence		
Scale 18	Selfless Service Orientation		
Dimension 2	**Technical Knowledge and Skill/Problem Solving**	.84	.92
Scale 02	Common Task Knowledge and Skill		
Scale 07	Level of Effort and Initiative on the Job		
Scale 24	Knowledge of the Inter-Relatedness of Units		
Scale 25	Management/Coordination of Multiple Battlefield Functions		
Scale 26	Problem-Solving/Decision Making Skill		
Dimension 3	**Adhering to Rules**	.79	.80
Scale 04	Safety Consciousness		
Scale 11	Adherence to Regulations, Policies, and Procedures		
Scale 12	Level of Integrity and Discipline on the Job		
Dimension 4	**Leadership**	.87	.90
Scale 16	Team Leadership		
Scale 19	Motivating, Leading, and Supporting Individual Subordinates		
Scale 20	Directing, Monitoring, and Supervising Individual Subordinates		
Scale 22	Concern for Soldier Quality Of Life		
Scale 23	Training Others		
Dimension 5	**Getting Along with Others (Interpersonal Skill)**	.64	.76
Scale 08	Adaptability		
Scale 15	Relating to and Supporting Peers		
Scale 17	Cultural Tolerance		
Dimension 6	**Information Management/Writing**	.67	.64
Scale 03	Computer Skills		
Scale 05	Writing Skill		
Scale 27	Information Management		
Dimension 7	**Oral Communication**	--	.58
Scale 06	Oral Communication		
No Dimension			
Scale 01	MOS/Occupation-Specific Knowledge and Skill		
Scale 09	General Self-Management Skill		
Scale 21	Modeling Effective Performance		

Table 2.6. Interrater Reliability for the Observed Performance Rating Scales

Scale	Interrater Reliability (single rater)	
	r	*ICC*
1. MOS/Occupation-Specific Knowledge and Skill	.44	.44
2. Common Task Knowledge and Skill	.25	.24
3. Computer Skills	.39	.39
4. Safety Consciousness	-.05	-.05
5. Writing Skill	.05	.05
6. Oral Communication Skill	.25	.25
7. Level of Effort and Initiative on the Job	.14	.14
8. Adaptability	.04	.04
9. General Self-Management Skill	-.08	-.08
10. Self-Directed Learning Skill	.11	.11
11. Adherence to Regulations, Policies, and Procedures	.34	.34
12. Level of Integrity and Discipline on the Job	.07	.07
13. Physical Fitness	.60	.58
14. Military Presence	.41	.41
15. Relating to and Supporting Peers	.06	.06
16. Team Leadership	.24	.24
17. Cultural Tolerance	-.12	-.12
18. Selfless Service Orientation	.24	.24
19. Motivating, Leading, and Supporting Individual Subordinates	.18	.18
20. Directing, Monitoring, and Supervising Individual Subordinates	.13	.13
21. Modeling Effective Performance	.31	.31
22. Concern for Soldier Quality of Life	.04	.04
23. Training Others	.36	.34
24. Knowledge of the Inter-Relatedness of Units	.21	.20
25. Management and Coordination of Multiple Battlefield Functions	.05	.05
26. Problem-Solving/Decision Making Skill	.13	.13
27. Information Management	.17	.17
Overall Effectiveness	.21	.21
Senior NCO Potential	.34	.34
Composite	.34	.34
Dimension 1: Fosters Improvement of Self and Others	.41	.41
Dimension 2: Technical Knowledge and Skill/Problem Solving	.38	.38
Dimension 3: Adhering to Rules	.22	.22
Dimension 4: Leadership	.32	.31
Dimension 5: Getting Along with Others (Interpersonal Skill)	.08	.08
Dimension 6: Information Management/Writing	.24	.24
Dimension 7: Oral Communication	.25	.25

Note. n = 37-72. For each soldier, one supervisor was assigned to the first rater group and another was assigned to the second rater group. Reliabilities were estimated by computing the correlation between the two groups of ratings for each scale, composite, and dimension (r) and the intraclass correlation coefficients (*ICCs*; Shrout & Fleiss, 1979, *ICC*[3,1] assessing consistency across a fixed set of raters).

To facilitate the task of evaluating soldier performance, some rating scales were consolidated based on the a priori model for performance components, the exploratory factor analysis results, and discussions among HumRRO and ARI project team members. As a result, five consolidated scales were developed to replace their associated individual scales (see Table 2.7). The anchor descriptions of the combined scales are composites of the individual scale anchors. These consolidated descriptions lose relatively little information (particularly given the high correlation between the overall composite and the single-item overall effectiveness rating) and preserve the future-oriented content of the scales. Thus, the validation version of the Observed Performance Rating Scales instrument is considerably shorter than the field test version, with a total of 21 scales (i.e., 19 specific performance scales, 1 overall effectiveness scale, and 1 senior NCO potential scale) instead of 29. This version of the Observed Performance Rating Scales is shown in Appendix A.

Table 2.7. Combined Observed Performance Rating Scale Items

Combined Scale	Original Individual Scales
• Self-Management and Self-Directed Learning Skill	• General Self-Management Skill • Self-Directed Learning Skill
• Demonstrated Integrity, Discipline, and Adherence to Army Procedures	• Safety Consciousness • Adherence to Regulations, Policies, and Procedures • Level of Integrity and Discipline on the Job
• Acting as a Role Model	• Physical Fitness • Military Presence • Modeling Effective Performance
• Leadership Skill	• Team Leadership • Motivating, Leading, and Supporting Individual Subordinates • Directing, Monitoring, and Supervising Individual Subordinates
• Coordination of Multiple Units and Battlefield Functions	• Knowledge of the Inter-Relatedness of Units • Management/Coordination of Multiple Battlefield Functions

Operational Implementation Options and Issues

After the research is completed, the Observed Performance Rating Scales (or a modification of this instrument) could be used for development-oriented performance appraisals. The instrument could also be used to collect diagnostic feedback information prior to attendance at NCO training courses (the Primary Leadership Development Course [PLDC] and/or Basic NCO Course [BNCOC]). If used this way, it would be best to collect ratings from a broad range of raters (supervisors, peers, subordinates, self), and instructors would need training on how to effectively counsel/train students based on this type of feedback. Whereas this type of instrument could be used to evaluate performance in the field, it should also be possible to develop an abbreviated version of the scales to collect training performance ratings from instructors and fellow students. The validation results are expected to inform the process of creating an abbreviated version of the Observed Performance Rating Scales.

Expected Future Performance Rating Scales

Army conditions are expected to evolve significantly over the next two decades resulting in changing NCO job requirements. The Expected Future Performance Rating Scales instrument is specifically designed to obtain supervisors' predictions of soldiers' performance in anticipated future Army conditions. The concept is based loosely on the Project A Combat Performance Prediction Scales (J. Campbell & Knapp, 2001). Because the focus of the NCO21 project is to improve the NCO promotion system in the future Army, it was necessary to develop a criterion instrument focusing on performance in future Army conditions. A primary purpose of the field test administration of this instrument was to assess the feasibility of the concept.

Instrument Development Process

The Expected Future Performance Rating Scales, like the Observed Performance Rating Scales, were designed to be suitable for all Army NCO jobs. Unlike the Observed Performance Rating Scales, however, the Expected Future Rating Scales were intended to be based on *projected* performance effectiveness, rather than *actual* observed performance. Initially, project staff conceived a format similar to that of the observed performance ratings using the same anchor descriptions—but with a different rating scale that assessed *expected* performance. This concept was rejected, however, due to the expected high correlations attributable to method variance that would result if a similar format were selected.

Measurement Method

Two primary concepts that would minimize common method bias were identified as possibilities for assessing future performance: (a) rate overall expected performance based on one or two scenarios that describe projected future Army conditions, and (b) rate expected performance based on several shorter scenarios, each targeted toward one or more projected future conditions. Project staff selected the latter option for three main reasons. First, the use of several targeted rating scales was anticipated to be more conducive to interpreting correlations between observed performance and expected future performance for a particular dimension. Second, it was expected that supervisors would more likely read several short, targeted scenarios rather than one long scenario. Finally, because most individuals tend to perform well in some areas and less well in other areas, using several scenarios was expected to differentiate expected performance on targeted areas better than one all-encompassing scenario. The use of one general scenario could potentially lead to a greater central tendency in responses, and perhaps a less accurate representation of expected performance.

Type of Rating Scale

Similar to the Observed Performance Rating Scales, a 7-point rating scale was developed with anchor levels at low (1-2), moderate (3-5), and high (6-7) performance. Unlike the other scales, however, the Expected Future Performance Rating Scales were worded in a general manner to reflect the likelihood that the soldier will meet or exceed NCO standards of performance in a given scenario. All scenarios used the same rating scale.

Predictions about future Army conditions and NCO job requirements focus on several themes. Table 2.8 displays nine major future conditions identified in Phase II of the project (Ford et al., 2000). Project staff involved in Phase II drafted two scenarios for the prototype instrument that primarily reflected the first three future conditions listed on Table 2.8 (Scenario 1 focused on self-direction and self-management whereas Scenario 2 focused on the use of computers and digitized equipment). The prototype scenarios were one-half to three-fourths of a page long.

Table 2.8. Anticipated Conditions in the 21st-century Army

- Greater requirement for self-direction
- Greater need to process information from multiple/digitized sources
- Greater requirement to use computers and/or computerized equipment
- Requirement for a broader range of technical skills
- Need for broad leadership skills will occur earlier (i.e., lower down) in the promotional ladder (e.g., today's E7 or E6 is tomorrow's E5)
- Greater need to know how a wide range of units and systems are interrelated
- Capability to manage multiple functions during an operation (i.e., future NCOs will do what today's captains and majors do)
- Greater physical and mental stamina
- More adaptability to new and varied missions

Pilot Testing the Prototype Instrument

The prototype Expected Future Performance Rating Scales instrument was administered to two pilot test samples (n = 88). The test administrator provided a brief overview of the purpose of the instrument following completion of the Observed Performance Rating Scales. NCOs were instructed to rate the same two E5 or E6 soldiers whom they rated using the Observed Performance Rating Scales. The raters read the two scenarios and projected how effectively their soldiers would perform under the conditions described. After making their ratings, the supervisors were asked to assess the extent to which they felt confident with their ratings of each soldier. The confidence scale ranged from 1 = not at all confident, to 5 = very confident.

The average ratings for the two scenarios were somewhat lower for the first pilot test sample than for the second (4.6 vs. 5.4 on Scenario 1 and 4.5 vs. 5.3 on Scenario 2 on a 7-point scale). The average confidence rating was 4.0 (on a 5-point scale) in the first pilot test and 4.3 in the second. This suggests that the supervisors were comfortable making projected ratings of future performance for the given scenarios.

The raters were asked to give feedback on the prototype instrument and they provided some suggested revisions to the two scenarios. They were also asked if the format was appropriate and if an alternative format (i.e., if they had a description of the current NCO requirements in addition to the description of the expected future NCO requirements) would be more effective. Most raters (59%) suggested that the instrument would be useful in the current format. Given that confidence ratings were generally high and a majority of the individuals were comfortable with the current format, no structural changes to the overall format were made.

Preparation for the Field Test

In preparation for the field test, project staff developed four additional scenarios (for a total of six) that targeted the remaining expected future Army conditions in Table 2.8. Because these scenarios would not be pilot tested, they were reviewed and revised multiple times based on input from ARI and HumRRO staff. The most significant modifications involved minimizing the degree of content overlap among the six scenarios. This shortened some of the scenarios, such that the amount of reading required was reduced to one-third to two-thirds of a page.

As with the Observed Performance Rating Scales, the Expected Future Performance Rating Scales were modified to allow for ratings of up to five soldiers per booklet. Several modifications were made to the instrument and to the rater training to facilitate this process. Finally, the middle anchor of the rating scales was revised to enhance its clarity. The Expected Future Performance Rating Scales are shown in Appendix B.

Field Test Administration

The Expected Future Performance Rating Scales instrument was administered to supervisors at three Army installations. Participants were given a brief verbal overview of the rating scales and their purpose. The supervisors were asked to read the six scenarios and provide ratings of expected performance effectiveness of their soldiers in those predicted future conditions. On average, administration lasted approximately 20 minutes, depending upon the number of soldiers to be rated. After completing their ratings, supervisors were asked to provide confidence ratings on a 7-point scale (1 = not at all confident, 7 = very confident) of the accuracy of their projected ratings.

Field Test Results

The field test administration of the Expected Future Performance Rating Scales yielded six item-level scores (ranging from 1 to 7) for each soldier from one or two supervisors. Higher scores indicate higher expected performance effectiveness in the specified future Army condition(s). The field test also provided scores on the supervisors' confidence in the ratings they provided for each soldier.

Descriptive Statistics

A total of 210 soldiers were rated by their supervisors on this instrument; 71 of these soldiers were rated by more than one supervisor. The means of the expected future performance ratings ranged between 4.80 and 5.10 (see Table 2.9); these means were slightly lower than those from the Observed Performance Rating Scales. There was a reasonable amount of spread in the scores—more than was found with the observed performance ratings. A composite score was computed by summing the mean ratings across all six scenarios; the mean composite score was 4.90 (SD = 1.07). The mean confidence rating was 5.72 (SD = 1.05) on a 7-point scale suggesting the supervisors were confident in their expectations of their soldiers' performance in future Army conditions.

Table 2.9. Descriptive Statistics for the Expected Performance Rating Scales

Scenario/Score	*M*	*SD*	*n*
1. Increased requirements for self-direction and self-management	4.83	1.28	210
2. Use of computers, computerized equipment, and digitized operations	4.97	1.23	209
3. Increased scope of technical skill requirements	5.10	1.23	209
4. Increased requirements for broader leadership skills at lower levels	4.88	1.28	210
5. Need to manage multiple operational functions and deal with inter-relatedness of units	4.80	1.25	209
6. Mental and physical adaptability and stamina	4.82	1.48	209
Composite Score	4.90	1.07	210

Subgroup Differences

Table 2.10 shows that there were no significant differences in the composite future performance score based on ratees' gender or race. However, there were differences between soldiers in combat and the other two MOS types, such that soldiers in combat MOS were rated higher. There was also the expected difference in composite scores by grade; E6 soldiers were rated higher than E5 soldiers.

Table 2.10. Subgroup Differences in the Composite Expected Performance Rating Scale Score

Group	*n*	*M*	*SD*	Effect Size	*p*
Gender					
Female	22	5.25	0.87	0.37	.094
Male	181	4.84	1.10		
Race					
Black	58	4.87	1.22	-0.05	.894
White	122	4.92	1.02		
Pay Grade					
E6	74	5.19	1.13	0.44	.004
E5	136	4.74	1.01		
MOS Type					
Combat Support	62	4.78	1.05	-0.46	.020
Combat	51	5.24	0.99		
Combat Service Support	88	4.79	1.14	-0.45	.018
Combat	51	5.24	0.99		
Combat Service Support	88	4.79	1.14	0.01	.938
Combat Support	62	4.78	1.05		

Note. Subgroup differences in composite score. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* referent group. Referent groups (e.g., Whites) are listed second in each pair. Care should be exercised when interpreting the statistics because unequal cell sizes and interaction effects were not taken into account.

Scale Building

Inter-item correlations. The ratings on the six scenarios were correlated to assess their degree of relationship (see Table 2.11). These item-level scores were significantly correlated, r = .51-.70, $p < .05$. In addition, the Expected Future Performance Rating Scale composite score was highly correlated with the Observed Performance Rating Scale composite score, $r = .77$, $p <$ .001.

Table 2.11. Inter-Item Correlations Among Expected Performance Rating

Scenario/Score	1	2	3	4	5	6
1. Increased requirements for self-direction and self-management	--					
2. Use of computers, computerized equipment, and digitized operations	.59	--				
3. Increased scope of technical skill requirements	.63	.66	--			
4. Increased requirements for broader leadership skills at lower levels	.68	.61	.69	--		
5. Need to manage multiple operational functions and deal with inter-relatedness of units	.64	.51	.63	.66	--	
6. Mental and physical adaptability and stamina	.70	.56	.62	.63	.66	--
Composite Score	.85	.78	.84	.85	.82	.84

Note. n = 208-209. All correlations are significant at $p < .01$.

Factor analysis. An exploratory factor analysis (EFA) was performed to determine if the Expected Future Performance Ratings measured more than one construct. The EFA used a maximum likelihood extraction with oblique rotation, specifying two, three, and four factors. The results for all three analyses showed high correlations among the factors and lacked simple structure. These results provided evidence for using a single composite expected score to summarize the data, rather than using the individual scenario scales.

An additional EFA was performed, including the ratings for the observed and future performance scores; this was conducted to determine if perhaps a "future performance" factor would emerge. The analysis was done two ways: (a) including all scores from both instruments, and (b) excluding the three scores from the observed performance ratings that did not have clear loadings in the observed performance factor analysis. Unfortunately, neither analysis could achieve simple structure; thus, the hypothesized model was not found.

Reliability Estimates

The internal consistency reliability (Cronbach's alpha) was computed for the composite expected future performance score. Similar to the observed performance internal consistency estimate, the reliability was high (α= .91). Inter-rater reliability was calculated using the same

methods as the observed performance ratings. These interrater reliability estimates were low, ranging from -.01 to .20 (see Table 2.12).

Table 2.12. Interrater Reliability for the Expected Future Performance Rating Scales

Scenario/Score	Interrater Reliability (single rater)	
	r	*ICC*
1. Increased requirements for self-direction and self-management	.12	.12
2. Use of computers, computerized equipment, and digitized operations	.10	.10
3. Increased scope of technical skill requirements	.01	-.01
4. Broader leadership skills at lower levels	.08	.10
5. Manage multiple operational functions and deal with inter-relatedness of units	.20	.19
6. Mental and physical adaptability and stamina	.20	.19
Composite Score	.16	.16

Note. n = 69-71. For each soldier, one supervisor was assigned to the first rater group and another was assigned to the second rater group. Reliabilities were estimated by computing (a) the correlation between the two groups of ratings for each scale and composite (r) and (b) the intraclass correlation coefficients (*ICCs*; Shrout & Fleiss, 1979, *ICC*[3,1] assessing consistency across a fixed set of raters).

Summary

The results of the analyses for the Expected Future Performance Rating Scales suggest that a composite score should be used to summarize the data. Although the instrument yielded high internal consistency, the interrater reliability estimates were generally low. No changes were made for the validation data collection.

As with the Observed Performance Rating Scales, the future-oriented scales may prove useful as an NCO professional development tool once the NCO21 research program has been completed.

CHAPTER 3: SITUATIONAL JUDGMENT TEST

Background

Situational judgment tests (SJTs) assess the effectiveness of examinees' judgments about the appropriate courses of action in various job-related scenarios. For the Army, the scenarios are usually supervisory situations. SJTs have been used since the 1920s, but they have become popular during the last 10 years. Two possible reasons for this surge in popularity are their demonstrated criterion-related validity and their intuitive appeal (i.e., the content of the test appears to be clearly relevant to the job). Meta-analysis has demonstrated the validity of SJTs for predicting job performance. The set of SJT studies analyzed by McDaniel, Morgeson, Finnegan, Campion, and Braverman (2001) yielded a correlation of .36 (correcting for criterion unreliability).[2]

Two SJTs were developed for the NCO21 project. The main test—the *SJT*—comprises items measuring the following eight NCO21 KSAs:

- Directing, Monitoring, and Supervising Individual Subordinates
- Training Others
- Team Leadership
- Concern for Soldiers' Quality of Life
- Cultural Tolerance
- Motivating, Leading, and Supporting Individual Subordinates
- Relating to and Supporting Peers
- Problem-Solving/Decision Making Skill

These KSAs were selected based on the extent to which (a) they were identified as measurable by the SJT and (b) the SJT would, in combination with other measures, provide adequate coverage of the KSAs identified as critical in Phase II of the NCO21 research program.

A supplemental test, the *SJT-X*, comprises items measuring Knowledge of Inter-Relatedness of Units. The SJT-X is separate from the SJT for two reasons: (a) its development process differed from the SJT, and (b) the items in the SJT-X contain lengthy scenarios—some requiring two pages of text. (In contrast, SJT scenarios are typically about three sentences long.) Unlike the SJT, which was finalized after analyzing field test data, the SJT-X requires additional data before the items and scoring algorithm can be finalized. The validation data collection will yield one potential source of these data.

[2] Ninety-one of 99 validity coefficients used supervisory ratings or rankings as the criterion, the remaining eight validity estimates used production data.

SJT Development

The SJT development process involved four major steps:

- Select potential items (i.e., scenarios and response options) from existing SJTs, and develop new items,
- Develop a scoring key,
- Prepare and administer a field test version of the SJT, and
- Select response options, items, and scoring algorithm for the final test (i.e., the version to be used in the validation effort).

Table 3.1 presents a summary of the SJT and SJT-X development activities.

The final step in the development process involved several interdependent goals. The choice of the "best" items and corresponding response options depends directly on the scoring algorithm applied. Similarly, selecting an optimal scoring algorithm necessarily depends on the items being scored. The interdependence of these selection decisions necessitated a complex automated analytic approach to select the optimal set of items, response options, and scoring algorithm. The procedures used to address this problem are described later in this chapter.

Table 3.1. SJT and SJT-X Development Activities by Location

Activity	Prior Studies	Fort Campbell	Fort Bragg	Fort Riley	USASMA	Fort Knox	Fort Leonard Wood	HumRRO
Collect old SJT items	x							
Collect old critical incidents	x							
Write critical incidents		x						
Write scenarios			x	x				
Write response options				x	x			
Edit scenarios				x				x
Edit response options								x
Rate effectiveness of response options					x	x	x	
Write SJT-X items								x
Rate effectiveness of SJT-X items						x	x	

Note. Except for HumRRO, the locations are listed in order of when their data collection took place. Each activity done at HumRRO took place at more than one time.

Item Generation

A literature review identified three SJT instruments of potential use to the NCO21 project, each of which was developed for the U.S. Army:

- The SJT developed for Project A (J. Campbell & Knapp, 2001),
- The Army Leadership Questionnaire (ALQ) developed for the ECQUIP project (Peterson et al., 1997), and
- The Platoon Leader Questionnaire (PLQ) developed by Hedlund et al. (1999).

These tests were reviewed to identify items that appeared to measure any of the target NCO21 KSAs. Although none of the dimensions covered by the ALQ, Project A SJT, or PLQ were the same as those we wanted to cover in the NCO21 SJT, all items from these instruments were considered for inclusion in the NCO21 SJT. The content of each item was examined for relevance to the target NCO21 KSAs. Several items from the ECQUIP and Project A SJTs did appear to be relevant. HumRRO staff members with Army experience reviewed and modified these items to improve their currency, accuracy, realism, and clarity. No PLQ items were retained because they either failed to relate to any of the NCO21 KSAs or were otherwise unsuited for our needs (for example, most of the PLQ items described situations more familiar to officers than to NCOs).

Although many useful items were obtained from the ECQUIP and Project A SJTs, many items remained to be developed to cover the target KSAs adequately. Two sources of SJT item scenarios were used. First, approximately 3,000 critical incidents from the ECQUIP (Peterson et al., 1997) and Special Forces (Russell et al., 1995) projects were content analyzed. Unfortunately, only a handful of incidents relating to NCO21 KSAs had enough detail to support development of new SJT items; most of these were related to Cultural Tolerance and involved getting along with foreign nationals during deployments. Critical incidents recorded during the development of the ALQ were also examined. The few ALQ items and incidents that were related to an NCO21 dimension were not relevant for lower grades or gave too little description of the situation to be useful.

Second, scenarios for new items were collected from senior NCOs at two Army sites and critical incidents were collected at one Army site. The NCOs were given instructions on how to write situation descriptions (or critical incidents). They were shown a sample situation description for each target KSA. The instructions were modified between each of the three data collections in an effort to improve the usefulness of the situation descriptions that were being produced. Thus, the last data collection yielded much better descriptions than the first.

HumRRO staff evaluated the situation descriptions and discarded those judged unable to support SJT items. The remaining descriptions served as the basis for new SJT item scenarios. They were edited for grammar, accuracy, realism, richness, clarity, and other aspects that would likely make the item a better measure of the target KSA. To the extent possible, NCOs also reviewed the scenarios (and response options) before the scoring key ratings were collected.

At two data collection sites, Fort Riley and the USASMA, NCOs wrote descriptions of what action should be taken in each situation. HumRRO staff members edited these responses.

Responses that were largely redundant were deleted. Remaining responses represented the draft response options for the scenarios.

Scoring Key Development

To develop a scoring key for the SJT, SME ratings of the effectiveness of each response option were needed for each of 105 draft SJT items. These ratings were obtained from 72 Sergeants Major (i.e., E9s) at USASMA. Because of an administrative problem, plans to equally divide the items across the SMEs could not be followed. Therefore, some items were rated by as few as 13 SMEs and others by as many as 33 SMEs.

After the SME rating session at USASMA, some SJT items were substantially revised (scenarios and response options) and some new response options were developed. Because there was no opportunity to obtain SME ratings for these items before the field test, SMEs at Fort Knox (7 E7 and 1 E8 NCOs) and Fort Leonard Wood (6 E7 and 2 E8 NCOs) rated these 31 items after the field test items had been finalized. The SMEs also rated five other items that fewer than 20 SMEs had rated at USASMA.

Preparation of Field Test SJT Forms

Having generated a set of scenarios, alternative responses to the scenarios, and experts' ratings of the effectiveness of the alternative responses, the next step was to decide which items to retain for the field test. The target for the final test was six items per KSA (i.e., 48 items). It was assumed that as many as one-third of the field test items would be dropped; therefore, nine items per KSA were needed in the field test.

Each item consisted of a scenario (i.e., stem) and four or more response options. For each item, the quality of the stem and response options was evaluated. These judgments of quality were based upon staff review and an analysis of the USASMA SME ratings. The analysis concerned two important qualities of ratings: interrater agreement and interrater reliability.

If SMEs disagree about the effectiveness of a response option (i.e., low interrater agreement), it is a poor option. Interrater agreement was assessed for each response option by computing the standard deviation among the raters.[3] Small standard deviations (i.e., low variability in the effectiveness ratings for a given response option) signify high interrater agreement. Most response options yielding low interrater agreement were discarded. Items with fewer than four acceptable response options were also dropped.

SMEs should agree not only about the absolute magnitude of their effectiveness ratings for a particular response option (i.e., low standard deviation), but also about the order of effectiveness of the set of response options for a particular item (i.e., high interrater reliability). Interrater reliability for the response option ratings (i.e., across all items) was computed using the intraclass correlation coefficient designated as *ICC(3,k)* by Shrout and Fleiss (1979), which is computationally identical to coefficient alpha. Coefficients were computed for three groups of

[3] Also computed for each response option was a frequency table that contained the number of SMEs who gave the option a particular rating (e.g., 1, 2, 3).

SMEs who rated the same set of items (recall that no SME rated all 105 draft items). Interrater reliability estimates for the three groups (which comprise 13, 23, and 33 raters, respectively) were .86, .92, and .94. Overall, the estimated interrater reliability of a single rater was .32—a respectable value for this statistic.

Given high interrater agreement and reliability, the effectiveness of the response options should vary within each item. The greater the spread in the options' effectiveness ratings, the greater the opportunity for the item to discriminate between effective and ineffective NCOs. Because soldiers were being asked to pick the best and worst responses, it was very important that the effectiveness of the best response option be substantially higher than that of the second-best response option. Similarly, it was important that the two worst (i.e., least effective) response options differ substantially in effectiveness. A draft item with two best or worst response options that did not differ in effectiveness could be retained, but one of the response options would have to be dropped in the final version.

Items were dropped if the highest rated two response options or lowest rated two response options differed by less than 0.7.[4] This rule would be less important if the final scoring algorithm (which had not yet been determined) used examinees' ratings of each response option rather than their choices for the best and/or worst responses. Therefore, a few items that violated this rule by a small amount were retained to fill out the number of items needed per KSA.

At this point, some KSAs still lacked items, and some items did not have enough response options. HumRRO staff with Army experience wrote a small number of additional scenarios and response options to address the shortfalls. At the end of this process, all KSAs but one contained at least nine items (there were eight items for Directing, Monitoring, and Supervising Individual Subordinates). Indeed, some KSAs had more than nine acceptable items. In these cases, one or more of the items sharing common themes (e.g., dealing with an overbearing supervisor) were discarded.

A total of 71 SJT items were administered in the field test. To make the SJT instrument a reasonable length, items were split into two forms (A and B). Items from two KSAs (Training Others; Directing, Monitoring, and Supervising Individual Subordinates) appeared on both forms. The items in the remaining KSAs appeared on either Form A or Form B. To the extent possible, similar KSAs were placed on the same form so that interscale correlations could be computed among similar scales.

Each test form included 44 items, instructions for completing the form, and two sample items. To answer each item, soldiers rated the effectiveness of each response option on a 7-point scale. They also picked the most and least effective response options (because some of the proposed scoring algorithms do not allow ties for the top-rated and bottom-rated response options). Because security is an issue with this potentially operational test, the SJT item shown in Figure 3.1 is only a sample item. The 7-point effectiveness scale is shown in Figure 3.2.

[4] The decision to use a difference of 0.7 had both a rational and practical basis: The difference had to be large enough to be meaningful (rational) but small enough to retain a sufficient number of draft response options (practical). The preferred difference value of 1.0 retained too few items.

In the field test, each SJT session was scheduled for 75 minutes (including oral instructions). The SJT was administered to E4, E5, and E6 soldiers. They had approximately 65 minutes to complete the 44 items.

One of your fellow soldiers feels like he doesn't have to pitch in and do the work that you were all told to do. What should you do?

Most/ Least	Effectiveness Rating (1–7)		
____	5	a.	Explain to the soldier that he is part of a team and needs to pull his weight.
L	2	b.	Report him to the NCO in charge.
M	5	c.	Find out why the soldier feels he doesn't need to pitch in.
____	3	d.	Keep out of it; this is something for the NCO in charge to notice and correct.
____	4	e.	Let him know that if he doesn't start doing his share you will report him to the NCO in charge.

Figure 3.1. Sample SJT item.

Effectiveness of the Action		
Ineffective action.	Moderately effective action.	Very effective action.
The action is likely to lead to a bad outcome.	The action is likely to lead to a passable or mixed outcome.	The action is likely to lead to a good outcome.
—— Low ——	**——— Moderate ———**	**—— High ——**
1 2	**3 4 5**	**6 7**

Figure 3.2. SJT response option effectiveness rating scale.

Plan for Selecting Response Options, Items, and Scoring Algorithm for the Final SJT

The primary purpose of the field test was to produce data that could be used to select the best set of response options for each item and the best set of items for the final version of the test. Analytical procedures were followed that first selected the best set of (a) four response options for each item after analyzing all possible sets of four options, and (b) five items for each KSA after several iterations of dropping the poorest remaining item in each KSA. These procedures were followed for six scoring algorithms (algorithms 1–6 described in Table 3.2). The following statistics were considered when determining the best set of response options: internal consistency reliability (Cronbach's alpha) for each scale, item-scale correlation, and item-total correlation. When determining which items to drop, these same criteria, plus breadth of construct coverage, were considered.

Selecting Response Options

Most field test items had more response options than needed (i.e., more than four). A good set of response options would evidence (a) high interrater agreement among the SMEs on the effectiveness of each response option, (b) a wide range of mean effectiveness values (i.e., mean rating by the SMEs) within a set of response options, (c) a keyed response having at least a moderate positive correlation with the scale score, and (d) distractors having negative correlations with the scale score (higher negative values are better).

Selecting Items

The field test intentionally had 50% more items than were needed for the final form of the test. The items for each of the eight SJT scales (i.e., KSAs) should have at least moderate positive item-scale correlations and high internal consistency reliability. The focus was on item-*scale* rather than item-*total* correlations for two reasons. First, it was important to try to develop an instrument that had useful—and, therefore, reliable—scores at the KSA level. For example, it would be useful to be able to say that a soldier was strong with regard to team leadership but weak in terms of peer relations. Second, it was important to cover as much of the domain as possible. If the scale scores proved unreliable, they would be less important and the total scores (i.e., including *all* items) would be more important. In this case, an item with a low item-scale correlation but a high item-total correlation might be retained.

Selecting a Scoring Algorithm

Six potential scoring algorithms were proposed for the SJT. These six algorithms can be put into two categories: those based on which options the candidates *select* as the most or least effective response and those based on the candidates' *ratings* of the effectiveness of the responses. Four of the six algorithms compared are actually combinations of simpler algorithms. There were six of these simpler algorithms. All algorithms are listed in Table 3.2. Algorithms 1–6 were evaluated. Algorithms a–f were used only during the computations of one or more of algorithms 1–6, as specified in the table.

Table 3.2. Alternative SJT Scoring Algorithms

Label	Short Description	Description of Algorithm (and Range of Possible Scores)
a.	1 if best = keyed best	Score is 1 if the candidate selects the keyed best response as the best response; score is 0 otherwise
b.	1 if worst = keyed worst	Score is 1 if the candidate selects the keyed worst response as the worst response; score is 0 otherwise
c.	–1 if best = keyed worst	Score is –1 if the candidate selects the keyed worst response as the best response; score is 0 otherwise
d.	–1 if worst = keyed best	Score is –1 if the candidate selects the keyed best response as the worst response; score is 0 otherwise
e.	Difference from keyed best	Score is the absolute value of the difference between the SME rating of the selected best response and the keyed best response (item score = 0 to 6). Lower scores reflect better performance.
f.	Difference from keyed worst	Score is the absolute value of the difference between the SME rating of the selected worst response and the keyed worst response (item score = 0 to 6). Lower scores reflect better performance.
1.	Best matches key (i.e., standard multiple-choice scoring)	Algorithm a only (item score = 0 or 1)
2.	Best matches key, worst matches key	Sum of algorithms a & b (item score = 0, 1, or 2)
3.	Best matches key, worst matches key, best reverse of key, worst reverse of key	Sum of algorithms a–d (item score = –2, –1, 0, 1, or 2)
4.	Difference from keyed best, difference from keyed worst	Sum of e & f (item score = 0 to 12; this range varies from item-to item because it depends on the SME mean ratings for the keyed best and worst responses). Lower scores reflect better performance.
5.	Difference from ratings key for all options	Using ratings: sum of differences between candidate's rating and SME mean rating for each response option (item score = 0 to 24; this range varies from item-to item because it depends on the SME mean ratings for each response option). Lower scores reflect better performance.
6.	Keyed value of best minus keyed value of worst.	Subtract SME mean rating of option selected as worst by candidate from SME mean rating of option selected as best by candidate (item score = –6 to +6).

Note. Keyed best response = the response option that received the highest rating of effectiveness from the SMEs, *keyed worst response* = the response option that received the lowest rating of effectiveness from the SMEs. An item's score using algorithms e, f, 4, 5, or 6 can be a non-integer; the other algorithms produce only integers.

Automating the Selection of Options, Items, and Scoring Algorithm

Selecting items, their corresponding response options, and a best scoring algorithm was complicated because the choice of one affects the results of the others. For example, item statistics depend on the choice of response options for the items, and the scoring algorithm in use affects the statistics for the response options and items. Further, one cannot focus on picking the best set of response options for each individual item and never revisit that decision. The options picked for one item can affect the options picked for another item. This is because the options picked are based partly on the item-scale correlation, and the scale score is affected by the options used for each item. In addition, the scoring algorithm used can affect the options picked.

Automated procedures that do not ignore or oversimplify this complexity were developed. The procedures selected response options, items, and a scoring algorithm on an empirical basis: they favored the combination with the best reliability, item-scale correlations, and item-total correlations. The final decisions were adjusted, however, based on human judgment. This "override" option was necessary to deal with situations that the automated procedures could not consider. For example, because an item's exclusion would greatly reduce the breadth of coverage of the target KSA, we might decide to retain it despite its attenuating effect on the scale's internal consistency reliability. Similarly, if two response options retained by the procedure were very similar in content, one might be discarded. These concerns were considered at the completion of the automated process.

The small number of alternative scoring algorithms (choosing one algorithm from among six) and response options per item (choosing four from as many as seven) permitted evaluation of all possible combinations of response options (n = 5,682).

SJT Field Test Results

Pre-Screening of Response Options

Response options with low interrater agreement among the SMEs were dropped before the automated procedures for selecting response options, items, and a scoring algorithm were executed. After this pre-screening process, some items had fewer than four response options. Many of these items were dropped; additional response options were written for others. After this prescreening process, 67 of the 71 items remained. Next, response options were selected for each item based on the item-scale and item-total[5] correlations.

Unfortunately, the scale score cannot be computed until the items in the scale are developed (i.e., the final set of response options are selected for each item) and chosen. To deal with this dilemma, a temporary scale score was computed based on all items in the scale and all response options for each item. After the first iteration, a new scale score was computed based on

[5] Item-total correlations were originally ignored while selecting response options. When it was discovered that the internal consistency reliabilities were somewhat low, the option-selection procedure was repeated while considering both item-total correlations and item-scale correlations. Item combinations (i.e., sets of four options) with the highest average of their item-scale and item-total correlations were selected. Thus, the item-total and item-scale correlations were given equal weight. The total score and scale scores were recomputed after each iteration.

the revised set of response options; this new scale score was used to compute the item-scale correlations and item-total correlations. This process of recomputing the scale scores and reselecting response options continued until the internal consistency reliability of the scale no longer improved. The steps in the entire automated procedure are given in Figure 3.3.

Each of the following steps was done separately for each of the six scoring algorithms.

Selecting Response Options

1. Compute a temporary scale score (i.e., KSA score) before dropping any of the items or response options (i.e., using all items and response options).
2. For each item, compute the item-scale and item-total correlations for all possible sets of four response options. (For each set of four response options, the keyed best response option was the one—among the four options in the set—with the highest mean effectiveness rating among the SMEs. In some cases, the option the *soldier* selected as the best response was not among the four options in the current set. In these cases, the option rated highest by the soldier was considered to be his/her selected option. If the soldier gave his/her two top responses the same rating, the scoring program randomly picked one of these as his/her selected response. The same procedures were followed for determining the keyed worst response and soldier's selected worst response.)
3. Compute the item-scale and item-total correlation for each item combination (i.e., for all possible sets of four response options). Compute a composite correlation by averaging the item-scale and item-total correlations. (The composite was computed because an item's correlations with its scale and with the total score were considered equally important.)
4. For each item, pick the set of four response options (i.e., potential item) with the highest composite correlation.
5. Compute a revised scale score based on the revised items. Repeat steps 2 and 3 until the scale reliability no longer increases. (As it turned out, no more than three iterations were needed for convergence. Some algorithms needed more iterations than others.)
6. Using the new scale and total scores, repeat steps 2–4. That is, determine whether different sets of response options are selected using the new scale and total scores. (As it turned out, the item combinations did not change.)

Selecting the Algorithm

7. Select the final algorithm based on scale reliability, total reliability, construct validity (based upon correlations with other measures), and practicality.

Selecting Items

8. In each scale, compute item-scale and item-total correlations for each item. (These are actually item-remainder correlations. That is, an item's score is not included in the scale score or the total score when it is correlated with the scale and total scores.)
9. In each scale, drop the item with the lowest item-scale correlation. (If the two lowest item-*scale* correlations differ by less than about .04, drop the item with the lower item-*total* correlation.)
10. Repeat steps 7–8 using the new scale scores until there are 5 items in each scale.
11. Drop items with very low item-total correlations. There need not be exactly the same number of items per scale. (In the end, there were only two items that might have been dropped at this stage. These items were retained for the validation data collection because the additional data might improve their apparent quality.)
12. Compute the reliability of the total SJT score.

Figure 3.3. Steps in the iterative automated procedure for evaluating SJT items, response options, and scoring algorithms.

Comparison of Scoring Algorithms

Table 3.3 shows the reliability of each of the six scoring algorithms computed. Only algorithms 1–6 were considered while choosing the response options and items for the final test form, but four of the algorithms are combinations of algorithms a–f. Neither Project A nor the ECQUIP project computed all six of the algorithms used here, but the results for the algorithms they did use are consistent with the current research. Table 3.4 compares the internal consistency reliabilities of three projects that used some of the NCO21 algorithms. Form B does not do as well as Form A or the SJTs in the other projects. Form A, however, is about as reliable as the ECQUIP SJT and more reliable than the Project A SJT or the Platoon Leader Questionnaire.

Table 3.3. Internal Consistency Reliability Estimates for Different Scoring Algorithms

				Form A					Form B				
	Scoring Algorithm	Form A	Form B	Tra	Sup	Peer	Cult	Mot	Tra	Sup	QoL	DM	Lead
	Number of Soldiers	181	211	206	221	219	215	206	227	233	222	230	231
	Number of Items	42	39	9	6	9	9	9	8	6	9	8	8
a	1 if best = keyed best												
b	1 if worst = keyed worst												
c	–1 if best = keyed worst												
d	–1 if worst = keyed best												
e	Difference between best and keyed best												
f	Difference between worst and keyed worst												
1	a only	.74	.56	.45	.27	.51	.54	.19	.17	.29	.13	.15	.34
2	Combine a & b	.80	.67	.52	.31	.57	.65	.29	.31	.26	.32	.09	.40
3	Combine a–d	.82	.68	.52	.34	.65	.69	.47	.26	.21	.35	.17	.38
4	Combine e & f	.87	.84	.58	.46	.73	.75	.63	.46	.36	.62	.58	.59
5	Difference from rating key for all options	.92	.92	.76	.62	.80	.78	.78	.68	.62	.74	.69	.75
6	Key for selected best – key for selected worst	.85	.72	.52	.43	.72	.76	.58	.43	.19	.43	.26	.46

Note. Because algorithms a–f were not considered on their own but were used only as part of algorithms 1–6, their reliabilities were not computed. Results are based on the best set of response options (after the first iteration) for each of the 67 items that passed pre-screening. The best response options were determined separately for each algorithm. Tra = Training Others; Sup = Directing, Monitoring, and Supervising Individual Subordinates; Peer = Relating to and Supporting Peers; Cult = Cultural Tolerance; Mot = Motivating, Leading, and Supporting Individual Subordinates; QoL = Concern for Soldiers' Quality of Life; DM = Problem Solving / Decision Making Skill, Lead = Team Leadership.

Table 3.4. Internal Consistency Reliability Estimates for Different Projects

Scoring Algorithm	NCO21 Form A	NCO21 Form B	ECQUIP	Project A	Platoon Leader Questionnaire
Number of Items	42	39	48	35	15
a 1 if best = keyed best	.74	.56	.58	.56	
b 1 if worst = keyed worst			.74	.48	
c −1 if best = keyed worst					
d −1 if worst = keyed best					
e Difference between best and keyed best			.75		
f Difference between worst and keyed worst			.84		
1 a only	.74	.56	.58	.56	
2 Combine a & b	.80	.67	.78		
3 Combine a–d	.82	.68	.82		
4 Combine e & f	.87	.84	.86		
5 Difference from rating key for all options	.92	.92			.69
6 Key for selected best – key for selected worst	.85	.72		.72	

Note. Because algorithms a–f were not considered on their own but were used only as part of algorithms 1–6, their reliabilities were not computed. NCO21 results are based on the best set of response options (after the first iteration) for each of the 67 items that passed pre-screening. The best response options were determined separately for each algorithm. Although the PLQ has only 15 items, it has almost twice as many response options per item as the other SJTs; because the PLQ uses only the scoring algorithm that uses all of the response option ratings, its reliability is a function of the number of response options rather than items.

Selection of Response Options, Items, and Scoring Algorithm

The automated process for evaluating various combinations of response options, items, and scoring algorithms was carried out as planned. Based on the process described above, and upon additional analyses and considerations, the following conclusions were made regarding the SJT form to be used in the validation data collection:

- There will be four response options for each item.
- It will use the Most Effective – Least Effective scoring algorithm from Project A (i.e., algorithm 6 in the current effort).
- It will contain 40 items: 5 items from each of the eight KSAs.
- Only a total composite score will be reported.

Having the same number of response options per item will make the test more consistent for examinees; it will also simplify completing, implementing, and maintaining the test. It was determined that four is a desirable and manageable number of options per item. For example, adding a fifth option per item would have lengthened testing time with little expected gain in

psychometric quality. The specific options to be retained on the final version of the SJT were identified through the automated evaluation program described previously.

The chosen algorithm computes the score for an item in the following way. The keyed rating (i.e., the SME mean rating) for the action selected by the soldier as *least* effective is subtracted from the keyed rating for the action selected by the soldier as the *most* effective. We have referred to this strategy as *algorithm 6*. Although two other algorithms (algorithms 4 and 5) had higher estimated reliabilities, they have two disadvantages. First, they require that soldiers rate the effectiveness of each action. Algorithm 6 requires the soldier to merely select the most and least effective responses. Thus, algorithm 6 requires less time to administer the test. In addition, using algorithms 4 or 5 would make it much more difficult to create an easy-to-understand machine scannable answer form, which is an issue for both our research and future implementation. Second, algorithms 4 and 5 did not correlate any better with the criteria. For example, they had non-significant negative correlations of -.01 and -.03 with the observed performance ratings whereas algorithm 6 had a small positive correlation of .05. All three algorithms had similar correlations with the interview, ASVAB, and AIM (see Chapter 7).

The decision to use five items per KSA is based upon the estimated reliabilities of the KSA scores and total scores for several possible numbers of items per KSA scale (i.e., 3 through 9). In general, the estimated reliability decreased only slightly as the number of items per scale dropped. In fact, when going from nine to five items per scale, the estimated reliability actually *increased* for six of the eight KSAs and for the total score for Form B. The KSA-level scores based on the best five items have estimated reliabilities ranging from .32 to .57 with a median of .43.

The 40 items to be retained for the final version of the SJT were selected based on their item-dimension score correlations, item-total score correlations, and content coverage. Despite our desire to derive scale (KSA) scores from the SJT, the factor analysis work described in the next section resulted in the decision to report only a total SJT score.

Dimensionality

Construct validity was examined by performing factor analyses and computing correlations among the SJT scales (examination of correlations with other measures is described in Chapter 7). The factor analyses examined item dimensionality and item loadings on the appropriate scales. The correlations with other measures determined whether the SJT scales were (a) related to other measures of the same constructs and (b) unrelated to measures of different constructs.

Factor analyses were performed on the final set of items to examine the dimensionality of the test. The initial goals of these analyses were to determine (a) whether scale scores should be reported, (b) the relationships between the scales, (c) the dimensions underlying the items, (d) how well the items fit the scales, and (e) whether replacing any items with discarded items would improve the fit of the items to the scales.

An exploratory factor analysis was performed for each SJT form using iterated principal factor extraction and oblique rotation. The number of factors was set to the number of scales (i.e., five factors for each form). The factor pattern matrix showed that items within the same scale did not typically load on the same factor (see Tables 3.5, 3.6). Each factor contained items

from several scales. In addition, each factor had only one or two high loadings. Finally, the highest loading was above .49 for only 13 of the 40 items.

The analysis was repeated while extracting the number of factors suggested by the scree plot of the eigenvalues and a parallel analysis. Separate analyses were done for the two SJT forms. The parallel analysis was done using Monte Carlo methods. For each form, 1,000 random datasets were generated with the same sample size and number of variables as the data used in the factor analysis. A scree plot was generated for each random dataset (the eigenvalues were based on the correlation matrix with squared multiple correlations in the diagonal). Each of these scree plots of random data was superimposed on top of the scree plot of the actual data. It was noted where the two scree plots crossed (i.e., after which eigenvalue number—or factor number).

Table 3.5. SJT Form A Factor Pattern Matrix

	Factor				
Item	1	2	3	4	5
Scale 1: Relating to Peers					
1	.05	.02	−.20	.24	**.57**
2	.17	.05	.20	**.46**	−.15
3	**.62**	.02	−.01	.14	.05
4	−.01	.11	−.09	**.28**	.12
5	−.01	**.38**	.14	−.05	.17
Scale 2: Intercultural Skill					
1	−.22	**.35**	−.20	.28	.16
2	**.28**	.18	.17	.15	−.11
3	−.03	**.59**	.12	−.05	−.06
4	.00	.17	−.00	**.46**	−.13
5	.04	**.65**	−.09	−.03	−.11
Scale 3: Motivating					
1	.03	−.01	.03	**.47**	−.06
2	.07	−.25	.02	**.51**	.13
3	−.19	**.27**	−.03	.16	.26
4	.26	**.46**	−.01	−.19	.15
5	.03	.11	**.62**	−.00	−.23
Scale 4: Training					
1	.10	−.11	−.09	**.67**	.06
2	.05	−.06	.02	−.09	**.71**
3	−.15	−.05	**.50**	−.19	.30
4	−.07	**.18**	.09	.17	−.02
5	.13	.06	−.03	−.14	**.39**
Scale 5: Supervising					
1	−.24	−.02	.13	.19	**.44**
2	.23	−.01	**.41**	−.21	.35
3	−.29	.20	.13	**.32**	−.04
4	−.04	−.16	**.49**	.34	.01
5	**.25**	.04	−.13	.06	.09

Note. n = 177. Each item's highest loading is boldfaced. Results based on final 40-item form (25 items in Form A).

For Form A, a four-factor solution was indicated by the parallel analysis and a sudden decrease in the eigenvalues after the fourth factor. For Form B, a seven-factor solution was indicated. Thus, an oblique factor analysis with four factors was performed on the Form A data. Similarly, a seven-factor oblique solution was generated for Form B.

Form A did not show simple structure. For example, only 7 of the 40 items had any factor loadings above 0.50 in the factor pattern matrix. Similarly, Form B did not show simple structure. Only 6 of the 40 items had any factor loadings in the pattern matrix above 0.49. Only the a priori five-factor solutions are shown here in Tables 3.5 and 3.6.

Table 3.6. SJT Form B Factor Pattern Matrix

	Factor				
Item	1	2	3	4	5
Scale 4: Training					
1	.16	−.02	**.19**	.18	−.00
2	−.10	−.17	**.48**	−.06	.02
3	**.43**	−.05	−.13	.07	.02
4	**.30**	.03	−.03	−.01	.11
5	−.06	**.18**	−.10	.02	.17
Scale 5: Supervising					
1	−.07	−.03	**.23**	.10	.23
2	−.01	−.16	.24	−.03	**.55**
3	.23	−.07	**.39**	−.04	−.06
4	−.09	.12	**.15**	.13	.05
5	**−.32**	.25	.07	.10	−.07
Scale 6: Concern for Quality of Life					
1	.09	.11	−.07	.19	**.19**
2	.01	**.73**	−.05	.02	.01
3	.08	.21	**.42**	−.35	−.01
4	.21	.27	.26	.01	**−.33**
5	.04	.18	−.12	−.17	**.55**
Scale 7: Problem Solving					
1	**.56**	.06	.06	−.01	−.01
2	.09	.01	−.08	**.48**	.13
3	**.19**	.15	.09	−.17	.00
4	−.03	−.22	**.43**	.15	.07
5	−.04	.07	.05	**.56**	−.19
Scale 8: Team Leadership					
1	.19	.05	**.24**	.23	.06
2	−.16	.21	**.32**	.07	−.07
3	−.11	**.26**	.17	−.02	.26
4	.10	.03	−.13	.13	**.43**
5	−.15	−.02	.16	.13	**.18**

Note. n = 196. Each item's highest loading is boldfaced; although some loadings in a row appear to be equal, they do differ at the third decimal place. Results based on final 40-item form (25 items in Form B).

The correlations among the scales were computed for each form. For Form A, the correlations ranged from .35 to .56 with a median of .48 (see Table 3.7). For Form B, they ranged from .24 to .36 with a median of .28 (see Table 3.8). The lower correlations for Form B are probably due to the lower reliabilities of its scales. The correlations cannot be accurately corrected for attenuation—to determine the correlations among the constructs underlying the scale scores—because the appropriate reliability estimates (test-retest with delay) could not be computed.

Table 3.7. Correlations Among Scales: SJT Form A

Scale	Relationships	Cultural Skill	Motivating	Training
Relationships				
Cultural Skill	.51			
Motivating	.56	.47		
Training	.48	.37	.39	
Supervising	.49	.35	.48	.42

Note. Results based on final 40-item form (25 items in Form A). Sample sizes range from 225–241. All values are greater than .19 and thus statistically significant at $p < .01$.

Table 3.8. Correlations Among Scales: SJT Form B

Scale	Training	Supervising	Concern for QL	Problem Solving
Training				
Supervising	.25			
Concern for QL	.28	.26		
Problem Solving	.36	.24	.29	
Team Leader	.28	.36	.26	.34

Note. Results based on final 40-item form (25 items in Form B). Sample sizes range from 234–245. All values are greater than .19 and thus statistically significant at $p < .01$.

Descriptive Statistics and Reliability Estimates

Table 3.9 shows descriptive statistics and reliability estimates for the final test. The estimated internal consistency reliability of a total score, assuming a 40-item test, was .84. It was computed by using the Spearman-Brown prophecy formula to estimate alpha for a hypothetical 40-item version of each form based on the computed alpha of each 25-item form. Then the average of the adjusted alphas for Form A and Form B was computed. Before averaging the two alphas, *r*-to-*z* transformations were performed; a *z*-to-*r* transformation was performed after averaging. The internal consistency reliability of the eight scale scores ranged from .32 to .57 with a median of .43.

The finding that the internal consistency is the same for Form A (25 items) as for the entire test (40 items) might cause some confusion. One might expect the score for the entire 40-item test to have a higher internal consistency than the 25-item total score. In this case, however, the low internal consistency of Form B prevents the internal consistency of the 40-item test from being higher.

Internal consistency reliability computations tend to underestimate the reliability of a situational judgment test because they assume the items in a scale are measuring the same thing. However, not only do SJT *scales* tend to be multidimensional, but even individual *items* often measure different things depending upon the soldier's response. Thus, the test-retest correlation, with a delay of at least one month between tests, is a much better estimate of reliability for a SJT. Unfortunately, no soldiers in the current project took the test twice.

Table 3.9. Estimated Internal Consistency and Interrater Reliability Estimates for the Final SJT

Form	Scale	*n*	Mean	*SD*	Scale-total *r*	Coefficient alpha	Interrater Reliability for Various Numbers of Raters				*n* of raters[a]
							1	15	30	45	
A	Total Score	246	2.49	0.74		.84	.40	.91	.95	.97	15-17
B	Total Score	249	1.94	0.51		.67	.40	.91	.95	.97	15-17
AB	Total Score[b]					.84					
A	Training	228	2.54	1.09	.54	.46					
B	Training	240	2.66	0.80	.42	.25					
AB	Training	468	2.60	0.95		.37	.44	.92	.96	.97	15-33
A	Supervising, Monitoring	242	2.24	0.88	.58	.46					
B	Supervising, Monitoring	248	2.25	0.72	.43	.25					
AB	Supervising, Monitoring	490	2.25	0.80		.37	.62	.96	.98	.99	16-31
A	Peers	228	3.30	1.02	.67	.57	.66	.97	.98	.99	15-17
A	Intercultural Skill	234	2.41	0.84	.53	.55	.68	.97	.98	.99	15-32
A	Motivating	234	2.19	0.79	.61	.44	.44	.92	.96	.97	13-43
B	Concern for Soldiers' Quality of Life	248	1.87	0.71	.40	.44	.43	.92	.96	.97	22-30
B	Decision Making / Problem Solving	239	1.21	0.85	.46	.32	.37	.90	.95	.96	16-32
B	Team Leadership	249	1.75	0.75	.50	.42	.44	.92	.96	.97	16-47

Note. Interrater reliability is the reliability of the SMEs' ratings of the response options (in the final 40-item test). This is shown within each scale and for the total test. When computing the scale-total correlations, the item's score was excluded from the total and scale scores.

[a]Different SMEs rated different parts of the test. The range of values shows the minimum and maximum number of SMEs per item.

[b]The coefficient alpha reported here estimates alpha for a hypothetical 40-item version of each form based on the computed alpha of each 25-item form. The average of the adjusted alphas for Form A and Form B was computed. Before averaging the two alphas, *r*-to-*z* transformations were performed; a *z*-to-*r* transformation was performed after averaging.

Because the scoring key is based on judgments, it is important to assess the reliability of the SMEs. This was done by computing interrater reliability. Its computation and interpretation were complicated by two aspects of the design. First, most SMEs rated only about half of the items. Second, the SMEs were not simply split across two parts of the test. That is, we did not have one set of SMEs rating one half of the test and another set of SMEs rating the other half of the test. Rather, there was wide variation in the number of SMEs per item. Therefore, a custom approach was needed to estimate interrater reliability.

The goal was to assess interrater reliability of the response option ratings for each scale and for the total test for one rater and for *k* raters (where *k* is the actual number of raters). Because the actual number of raters varied from item-to-item, interrater agreement was estimated for 15, 30, and 45 raters. The three numbers represent, approximately, the minimum, median, and maximum number of raters among the items.

The mean correlation of response option ratings (mean was computed after doing an *r*-to-*z* transformation) for all possible pairs of SMEs estimated the reliability of judgments made by a single SME. Using the Spearman-Brown formula, the reliability of 1, 15, 30, and 45 raters was estimated.

Subgroup Analyses

Mean differences between groups were calculated to determine whether minority and gender groups performed differently on the SJT. For each form, *t*-tests were computed to compare the test scores of white and black soldiers, and to compare males with females. Tables 3.10 and 3.11 show the results of these analyses. According to the *t*-tests, there were no significant effects of gender or race upon the scores.

Table 3.10. Subgroup Differences in the SJT Form A Scores

Group	*n*	*M*	*SD*	Effect Size	*p*
Gender					
Female	43	2.61	0.68	0.18	.241
Male	196	2.48	0.74		
Race					
Black	73	2.42	0.76	–0.16	.275
White	136	2.54	0.74		
Pay Grade					
E5	103	2.67	0.60	0.56	<.001
E4	95	2.19	0.85		
E6	47	2.74	0.57	0.65	<.001
E4	95	2.19	0.85		
E6	47	2.74	0.57	0.12	.522
E5	103	2.67	0.60		
MOS Type					
Combat Support	69	2.44	0.79	0.09	.586
Combat	60	2.37	0.75		
Combat Service Support	108	2.58	0.70	.0.28	.087
Combat	60	2.37	0.75		
Combat Service Support	108	2.58	0.70	0.19	.268
Combat Support	69	2.45	0.68		

Note. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* referent group. Referent groups (e.g., Whites) are listed second in each pair. Care should be exercised when interpreting the statistics because unequal cell sizes and interaction effects were not taken into account. Results based on final 40-item form (25 items in Form A).

SJT scores were also compared by MOS type and pay grade (see Tables 3.10 and 3.11). Similar to the gender and race analyses, *d*-scores (effect sizes) were computed and *t*-tests performed. Only one significant difference was found among MOS categories: In Form B, soldiers in combat support MOS had a significantly higher mean score than did those in combat MOS; the difference was not significant in Form A.

Because E5 and E6 soldiers have more supervisory experience than E4 soldiers, one would expect them to do better on the SJT. The analyses show that soldiers at the E5 and E6 levels did, in fact, significantly outperform E4 soldiers. For example, E5 soldiers scored an average (across the two forms) of 0.57 standard deviations higher than E4 soldiers. There was no appreciable difference between the performance of E5 and E6 soldiers. This is not too surprising because both E5 and E6 soldiers should have a good bit of supervisory experience.

Table 3.11. Subgroup Differences in the SJT Form B Scores

Group	*n*	*M*	*SD*	Effect Size	*p*
Gender					
Female	29	1.75	0.75	–0.45	.140
Male	209	1.96	0.47		
Race					
Black	51	1.91	0.50	–0.13	.425
White	151	1.97	0.52		
Pay Grade					
E5	103	2.04	0.46	0.57	<.001
E4	94	1.73	0.54		
E6	49	2.17	0.34	0.81	<.001
E4	94	1.73	0.54		
E6	49	2.17	0.34	0.28	.065
E5	103	2.04	0.46		
MOS Type					
Combat Support	59	2.02	0.46	0.37	.040
Combat	61	1.84	0.48		
Combat Service Support	119	1.95	0.55	0.23	.173
Combat	61	1.84	0.48		
Combat Service Support	119	1.95	0.46	–0.15	.370
Combat Support	59	2.02	0.46		

Note. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* referent group. Referent groups (e.g., Whites) are listed second in each pair. Care should be exercised when interpreting the statistics because unequal cell sizes and interaction effects were not taken into account. Results based on final 40-item form (25 items in Form B).

SJT-X Development and Results

The purpose of the SJT-X is to measure Knowledge of the Inter-Relatedness of Units. No critical incidents or SJT items from previous studies assessed this KSA. Moreover, the NCOs who generated SJT scenarios for this project were unable to provide scenarios because this KSA addresses future functions most NCOs have not experienced. Another problem was that brief scenarios were unable to portray the complex situations in which issues associated with the inter-relatedness of units would arise. After many attempts, three items were developed using input from a Command Sergeant Major at Fort Riley and a retired Army officer on HumRRO's staff. Both of these individuals are knowledgeable about the expected future requirements regarding the target KSA.

The SJT-X scoring key was developed by administering the instrument to 16 NCOs at the E7 and E8 levels. The interrater reliability for the mean ratings of the 16 SMEs was .89. The estimated 1-rater reliability was .35.

The SJT-X was administered to 24 soldiers at one field test site (Fort Leonard Wood). Only E6 participants (a) were expected to understand the items and (b) had sufficient time in their test schedule to add the SJT. Because of the small sample size, all analyses of the SJT-X are only suggestive; no changes were made to the SJT-X based on the results. Table 3.12 shows the results of the item analysis of the SJT-X and reliability estimates. The internal consistency reliability (Cronbach's alpha) of the response option scores was computed within each item and for the test as a whole. Based on these results, the items appear to lack sufficient reliability.

Several options have negative correlations with the item scores. Thus, one would expect that dropping the options with the lowest correlations would increase reliability. This is confirmed in Table 3.13. The reliability estimates increase dramatically when the best set of four options (i.e., the set of four options that has the highest reliability) for each item are retained (see Table 3.13). Because of the small sample size, the values for the reliabilities in the population likely differ somewhat from these, but we can be somewhat confident that dropping options can improve the items substantially. [6]

The SJT-X items were re-scored using only the four best options from each item. A manual process was used to determine which options to drop. During each iteration, the option was dropped that maximized alpha for the set of remaining options. The correlations among the three SJT-X items scored in this fashion are shown in Table 3.14. Although the sample size is too small to draw any firm conclusions, the scores suggest that item 1 is unrelated to the other two items. In addition, the correlation between items 2 and 3 is only moderate. Thus, these three items appear to be measuring somewhat different things. They could be either measuring different aspects of Knowledge of Inter-Relatedness of Units or completely different constructs. If the three items are measuring different aspects of the KSA, then low correlations are expected. In fact, high inter-item correlations in such a short test would be undesirable: They would indicate that the test covers only a small portion of the construct domain.

Given the very small amount of data collected in the field test, the results are only suggestive and therefore were not used to revise the instrument. Moreover, unlike the SJT, examinees in the validation data collection will respond to items by rating the effectiveness of each response option (in addition to identifying the most and least effective responses). This method of responding will help to compensate for the test's small number of items. Concerns remain, however, about the extensive reading requirements of the items and the lack of opportunity soldiers (especially at lower grades) currently have to experience situations requiring knowledge of unit inter-relations. Therefore, as in the field test, current plans call for administering the SJT-X only to E6 soldiers in the validation data collection.

[6] We are not advocating choosing which response options to drop based on the sole goal of maximizing internal consistency. When selecting response options for the final SJT-X, breadth of construct coverage should also be considered. The analyses were done merely to see if internal consistency *might* be increased to an acceptable level when more data are collected.

Table 3.12. Item Analysis Statistics for All SJT-X Options

Item / Option	Coefficient Alpha	Option-Total Correlation	Option-Item Correlation	Interrater Reliability
Item 1	.35			.91
Option 1		.37	.18	
Option 2		.07	−.10	
Option 3		.15	.29	
Option 4		.34	.45	
Option 5		.26	.34	
Option 6		.24	−.13	
Option 7		.05	.08	
Item 2	−.73			.82
Option 1		−.05	−.51	
Option 2		.18	.02	
Option 3		−.00	−.00	
Option 4		−.28	.09	
Option 5		−.28	−.39	
Option 6		.12	.03	
Option 7		.10	−.40	
Item 3	.31			.92
Option 1		.23	−.03	
Option 2		.26	.42	
Option 3		.06	.20	
Option 4		−.05	−.19	
Option 5		.61	.49	
Option 6		.18	.10	
Option 7		−.41	−.41	
Option 8		−.06	.06	
Option 9		.07	.03	
Option 10		.41	.28	
Option 11		.13	.26	
Option 12		.46	.27	
All Items Combined	.43			.89

Note. n = 24. Interrater reliabilities are based on 16 raters. The option was removed from the total score before computing its option-total and option-item correlations.

Table 3.13. Item Analysis Statistics for the Best Set of SJT-X Options

Item / Option	Coefficient Alpha	Option-Total Correlation	Option-Item Correlation	Interrater Reliability
Item 1	.55			.78
Option 3		.–.06	.22	
Option 4		.21	.37	
Option 5		.08	.54	
Option 7		–.07	.25	
Item 2	.41			.95
Option 2		.19	.01	
Option 3		.37	.39	
Option 4		–.07	.23	
Option 6		.33	.32	
Item 3	.69			.97
Option 2		.14	.24	
Option 5		.49	.53	
Option 11		.19	.49	
Option 12		.75	.68	
All Items Combined	.54			.85

Note. n = 24. Interrater reliabilities are based on 16 raters. The option was removed from the total score before computing its option-total and option-item correlations.

Table 3.14. Correlations Among Revised SJT-X Items

	Item 1	Item 2	Item 3
Item 1		–.22	–.03
Item 2	–.22		.37
Item 3	–.03	.37	

Note. n = 23–24. No correlations are significant, but the sample size was so small that correlations must be greater than .40 to be significant at $p < .05$.

Summary

The final version of the SJT consists of 40 items that cover eight NCO21 KSAs. Although each item was selected to measure one of the eight KSAs, only a single overall score will be computed for the test. Separate scores for each KSA are not reported because they lack sufficient construct validity. This is not surprising because SJT test items typically tap more than one dimension, so most SJTs do not report dimension scores. When administered in the validation data collection, soldiers will be instructed to pick the most and least effective actions for each item. Tests will be scored using algorithm 6 (the *Most – Least Effectiveness* algorithm).

If the SJT is implemented as part of a promotion system, it would probably be easiest to manage if it were administered and scored via computer with delivery probably through a secure internet-based system. This would also be key to maintaining security, as the test items would be

kept under strict control, and it would be less burdensome on the Army to centralize administration of the instrument. It will also be necessary to develop alternate forms. It should be relatively easy to design an ongoing process for developing new items that would be neither unduly difficult nor expensive to implement. Of course, it would also be necessary to establish polices for when soldiers can take the SJT and how the scores would be used (e.g., meet a minimum cutoff or have the total points contribute to a total promotion point worksheet score). There would also be the need to establish operational procedures for administering the test, presumably on demand.

It would also be worth considering using the SJT or a variation of it for training and development. The items could be used for training in three ways. Before training, they could be used to assess training needs. During training, they could be used to illustrate what to do in various situations and provide practice in applying principles to situations. After training, they could be used to assess the effectiveness of the training. More sample items would become available as operational items are retired. Use of the SJT approach in training and development activities could also be a relatively transparent strategy to generate new items for the test if it is used operationally for promotion decisions.

Because soldiers at the E4 level do not typically have much supervisory experience, they are probably unfamiliar with some of the situations described in the SJT. However, based on the hypothesis that E4 soldiers will be given more supervisory responsibility in the future, we expect the SJT to eventually become a better predictor of performance.

The SJT-X consists of three items that measure Knowledge of Inter-Relatedness of Units. It has high interrater reliability and the limited data gathered thus far suggest that it can achieve sufficient reliability if the response options for the final test form are selected judiciously. This instrument is not well-suited to today's soldiers, but we expect it will become more appropriate for administration to soldiers in the target grades (i.e., E4 and E5) in the future.

CHAPTER 4: ARCHIVAL AND EXPERIENCE MEASURES

Overview

This chapter describes the development of two instruments designed to assess soldier work background (e.g., experiences, activities, and accomplishments) and archival information (e.g., test scores, commendations, awards, course credits). The Experience and Activities Record (ExAct) consists of self-report items designed to capture information about specific soldier experiences and activities that are typically not documented. The Personnel File Form–21 (PFF21) consists of self-report items designed to quickly and efficiently capture information that is normally documented and otherwise available in archival records.

Much of the initiative to develop assessments of experience and archival information stems from the previous success of similar measures in Project A. Multiple self-report instruments were developed during Project A to capture biodata (e.g., Assessment of Background and Life Experiences), archival information (e.g., Personnel File Form), and soldier experiences (e.g., Supervisory Experience Questionnaire). In that project, these instruments provided information that predicted soldier performance (see J. Campbell, 1987, for a review of Project A instrument development).

Although the ExAct and the PFF21 were developed simultaneously and influenced the development of one another to some degree, each instrument is discussed separately for ease of explanation.

Experience and Activities Record

Overview and Background

The ExAct is designed to assess the extent to which a soldier has engaged in specific activities or had particular experiences that may predict performance at the next grade. It is a reasonable presumption that soldiers who have engaged in more of these activities and have done so more often will perform at a higher level than those with less experience. That is, knowledge of a soldier's prior experiences should provide useful information for assessing his or her preparedness to perform similar activities in the future. This concept has been the basis for the development and use of accomplishment records (Hough, 1984) and biodata scales (Mael, 1994; Stokes & Toth, 1996). Personnel research has shown these types of scales are valid predictors of criteria such as leadership (Mael & Hirsch, 1993), managerial progress (Carlson, Scullen, Schmidt, Rothstein, & Erwin, 1999), performance ratings (Hough, 1984; McManus & Kelly, 1999; Mitchell, 1994; Vinchur, Schippmann, Switzer, & Roth, 1998), accidents (Hansen, 1989), and attrition (Laurence, 1990; Mael & Ashforth, 1995).

Instrument Development Process

A variety of sources were used to generate ExAct items. First, a focus group of 18 NCOs (E6-E9) provided feedback on the general concept and provided suggestions for item content. This information was used to generate a 44-item prototype measure that was pilot tested on a sample of 60 soldiers ($n_{E4} = 29$; $n_{E5} = 31$). To reduce the potential deleterious effects of response

distortion on self-report measures (see Zickar & Robie, 1999), items were limited to those that were historical, external (i.e., observable behaviors), and verifiable in principle. Items were further reviewed to ensure they (a) were related to at least one of the identified KSAs as relevant for 21st-century NCOs (see Table 1.3) and (b) reflected behaviors appropriate and reasonable for E4 or E5 soldiers to perform. Concurrently, a second focus group of 48 NCOs (E6-E9) provided reactions to the general concept and provided additional experience items related to specified KSAs (e.g., Computer Skills; Directing, Monitoring, and Supervising Individual Subordinates). Feedback from the pilot test and the second focus group resulted in 10 new items. The second prototype was administered to 43 soldiers ($n_{E5} = 2$ and $n_{E6} = 41$) representing 18 MOS. Following completion of the form, soldiers were asked (a) whether they thought this type of information was relevant for promotion decisions and why, (b) whether any items were confusing or vague, and (c) to generate of list of additional experiences or activities.

Overall, the results of the pilot test administrations showed variability within and across items, suggesting widespread response distortion was not occurring. Consistent with expectations, the frequency of self-reported experiences by E6 NCOs was slightly elevated compared to E5s, which in turn was elevated compared to E4s. Inspection of individual items revealed a few items with poor overall variance. Six items for which more than 75% of the soldiers within each grade reported never having the experience were deleted.

Ninety-nine soldiers ($n_{E4} = 26$, $n_{E5} = 32$, and $n_{E6} = 41$) responded to the follow-up question concerning whether soldier experiences should be assessed for promotion purposes. The 60 soldiers (61%) in favor of using experience information cited five main reasons why the ExAct would supplement the current system: (a) those who are more experienced and competent will receive promotions, (b) it is relevant to a soldier's actual duties and job experience, (c) it provides a better assessment of differences in leadership experience and skills, (d) it recognizes demonstrated effort and initiative, and (e) it is perceived as a more objective assessment of promotion potential. Of the 39 (39%) who did not favor the use of soldier experiences for promotion purposes, only 13 gave reasons that were specific to the ExAct. These reasons included beliefs that (a) there would be differential opportunity to have certain experiences, (b) the items do not reflect job competence, and (c) promotion should not be based on a soldier's prior job performance.

To assess the prototype items' relevance for promotion readiness for E4 and E5 soldiers, 34 NCOs (E7-E9) from Fort Riley and USASMA evaluated the items. In accordance with a priori expectations, most items were evaluated as slightly more relevant for E5 promotion readiness than for E4 readiness. Examination of item content showed items relating to leadership experience, personal effort, and formal training activities were generally seen as the most relevant. The only substantial grade-level differences appeared with the communication-related items (e.g., writing orders and reports, delivering briefings) with higher ratings for E5 than E4 promotions. Computer-related experiences were seen as being of marginal importance for both grades. Note, however, the soldiers were making the ratings based on current importance. It is likely these types of items will increase in importance in the future. A number of items were removed or reworded on the basis of the pilot test results, yielding the 46-item version used in the field test.

Field Test Administration

In part because the ExAct could potentially be scored using the rainforest empiricism approach (described in the next section), it was administered to E6, as well as E4 and E5, soldiers in the field test. To allow the required scoring for both predictor and criterion measurement purposes, two separate forms were developed. The version given to E4 soldiers asked each question only once, whereas the E5/E6 version asked soldiers to answer all questions twice – once based on their experience prior to their last promotion and once based on their experience while in their current grade.

Scoring Key Development

Discussion of Alternative Methods

Although a number of methods have been used to score biodata-type instruments, most have serious drawbacks that limit their utility. Empirical keying (a.k.a. "Dustbowl Empiricism") weights each alternative of each item based on its mean score on the criterion of interest. Although this method has been used successfully in some cases (e.g., for vocational interest inventories), it has been criticized as being atheoretical, leading to illogical scoring keys, low construct validity, and substantial shrinkage in cross-validations (Mael & Hirsch, 1993). Another common method, the rational approach, attempts to measure unitary constructs by combining items into homogeneous scales. This method relies on theory and is more robust to sample-specific fluctuations than is empirical keying. Biodata items typically reflect behaviors that draw on a heterogeneous collection of individual characteristics, however, so they cannot be easily linked to a single underlying construct. Thus, the rational approach may also be unsatisfactory because it (a) treats items as pure measures of single constructs even though they may be a function of multiple KSAs and (b) disallows the possibility the same behavior can be differentially related to various criteria (Mael & Hirsch, 1993). Another strategy involves creating scales based on factor analysis. Although this method allows items to differentially contribute to more than one factor, the factor-based scales are often psychologically uninterpretable. The primary advantage is a small number of scale scores can reflect primary components of common variance.

Mael (1991; Mael & Hirsch, 1993) proposed a variation of empirical keying, termed "rainforest empiricism," which relies on theoretical considerations in the choice of items and in keying decisions. With this method, a biodata scale is created for each criterion to be predicted. Items are initially keyed empirically; however, items are selected and keys are modified based on rational judgments. Using this approach, items are initially selected on a theoretical basis. Items that display poor overall variance are removed and item response alternatives with low endorsement rates are combined with the adjacent response alternative. After empirically keying the selected items, illogical key patterns are corrected. For instance, suppose the criterion means of the response alternatives for the item "*Total time spent in a leadership or supervisory position*" were 2.4 (Never), 3.2 (Less than 6 months), 3.8 (6 months to a year), 3.0 (1 year to 2 years), and 4.3 (More than 2 years). Using the strict empirical approach, the scoring key would be set such that having 1 to 2 years of experience is worse than having less than 6 months. In the rainforest empiricism approach, the key would be corrected to reflect a more rational scale (i.e., more experience receives increasingly higher key values). The rainforest empiricism approach retains the advantages of empirical keying while ameliorating some of its major drawbacks.

Similarly, while capturing some of the advantages of rational scaling, this approach does not suffer from the drawbacks associated with rational scale construction. On the downside, this approach requires large sample sizes with responses to the biodata instrument and criterion data. Because the scoring key is based on the average criterion score of those who responded the same to a given item (e.g., all those who responded "3" to item 4), the total sample size must be rather large so that each individual key value is based on a reasonable sample.

Although a "rainforest empiricism" keying strategy would be appropriate for the ExAct, the field test did not provide a sufficiently large sample to afford the use of this approach. Therefore, an empirically guided rational approach was used to score the ExAct. In this method, a mixture of rational judgment and empirical guidance is used to determine (a) the number of scales that should be derived and (b) which items should be retained and scored.

The ExAct items were developed to assess several KSAs. Item writers targeted at least 15 different KSAs during the course of instrument development, though the resulting items could be interpreted as only covering elements of 5 or 6 at most (e.g., Writing Skills, Computer Skills, Team Leadership). Because biographical items typically reflect multiple KSAs (in varying degrees), a total score for such an instrument is often used. A total score is inappropriate, however, if there are relatively independent dimensions clearly defined by specific items. A total score based on two orthogonal dimensions, for example, would have equivocal meaning because the same total score could reflect a variety of (unknown) combinations of scores on the two dimensions. This is further exacerbated by the extent to which the dimensions have differential relationships across criterion domains. To investigate this, principal components analysis can be used to determine if more than one primary component of variance is evident. Given the multicolinearity among the targeted KSAs and the multifaceted nature of the items, it was not expected the principal components would reflect the KSAs.

Item Analyses

Because the sample size did not allow for the rainforest keying method, separation of the "prior" and "current" responses was unnecessary with regard to the E5/E6 soldiers. This raised the question, then, of which responses to use for these soldiers. Later we discuss how the "timing" issue will be addressed in the validation effort. For purposes of the field test data analysis, however, an effort was made to eliminate the time distinction (prior vs. current) because the meaning of the distinction was ambiguous (e.g., "current" could span 2 months or 2 years). Moreover, the prior and current responses for most items correlated very highly. For the first 24 items, the time distinction was removed by adopting the highest frequency response (whether denoted as prior or current) as the scored response. The last 22 ExAct items asked respondents to indicate the total number of times they had experienced certain training or duty assignments (e.g., number of times on a combat mission). For these items, the soldiers' "prior" and "current" responses were summed. Examination of the raw responses indicated that most items had acceptable variability, with two or more response alternatives showing at least a 10% endorsement rate.

Scale Analyses

Prior to principal components analysis, all items were first standardized ($M = 0$, $SD = 1$) to place them on the same metric. For both E4 and E5 samples (the E6 respondents were not

included in these analyses), evaluation of the eigenvalues revealed two components were appropriate to describe the data. After rotation, a clear pattern emerged with all of the computer-related items (and primarily only those items) loading saliently on one component, and most of the remaining items loading primarily on the other. A few items were complex, loading saliently on both components, and a few failed to load on either component. Because of the relatively clean pattern of loadings for most items, it was decided that two scores would be appropriate: a Computer Experience score (Items 1-8) and a General Experience score consisting of all other items (Items 9-46). The Computer Experience score is a simple sum of items. Because items on the General Experience score had varying response options (ranging from 3-5 scale points), the items were standardized prior to summing.

To further evaluate the scores, corrected item-total correlations and item-deleted alpha estimates were computed and evaluated. Although an estimate of internal consistency may seem inappropriate for heterogeneous biodata items, it can be used because it does not require strict unidimensionality. Alpha estimates the proportion of test variance due to all common factors (both general and group) among the items (Cronbach, 1951). High alpha levels may suggest strong interrelatedness among the items, but a set of items can be interrelated and still be multidimensional (Cortina, 1993). Overall, all items are internally consistent with the scale scores. Item-deleted alphas did not indicate significant improvement in internal consistency with the removal of any of the items. For only a few items did removal result in a minimal increment in alpha (e.g., maximum observed increment = .02). Content analysis of these items suggested they were conceptually consistent with the scale. Therefore, all items were retained and scored. The alpha coefficients for Computer Experience were $\alpha = .86$ for the E4 sample, $\alpha = .82$ for the E5s, and $\alpha = .81$ for the E6 soldiers. Internal consistency estimates for General Experience were slightly higher ($\alpha = .91$ for E4s, .89 for E5s, and .83 for E6s). The correlation between the Computer Experience and General Experience scores was $r = .25$ for E4s, $r = .19$ for E5s, and $r = .16$ for E6s. These low correlations support the decision to use two scores.

Descriptive statistics for the scores are shown in Table 4.1. Recall the General Experience score was standardized. As expected, mean scores on both scales are higher with advancing pay grades. Subgroup analyses were conducted to examine mean differences based on gender, race, and MOS type. These analyses were conducted only for the grade levels for which the ExAct is targeted (E4 and E5). Differences by grade were also examined; these analyses included the E6 data. Group means, standard deviations, *p*-values, and effect sizes are shown in Table 4.1.

Looking first at E4 level gender differences, there is a difference for both the Computer Experience and the General Experience scores, although in opposite directions. Women have a higher mean Computer Experience score than men ($d = .57$), whereas men have a higher General Experience score than women ($d = -.49$). This pattern is repeated at the E5 level, although the effect sizes are smaller for the computer score ($d = .24$) and larger for the general score ($d = .75$). Examination of racial differences showed that, at both the E4 and E5 level, the means are very close for Computer Experience; blacks score just a bit lower on the General Experience Score. Differences based on type of MOS show a pattern consistent with expectations. Soldiers in combat MOS score, on average, higher on the General Experience score than other MOS, and score lower (at the E4 level) on Computer Experience than other MOS. Finally, the expected escalation in scores as grade increases was observed in both ExAct scores.

Table 4.1. Descriptive Statistics for the Experience and Activities Record Scores

	Computer Experience Score					General Experience Score				
Group	*n*	*M*	*SD*	Effect Size	*p*	*n*	*M*	*SD*	Effect Size	*p*
E4										
Gender										
Female	34	30.29	9.44	0.57	.003	34	-26.04	12.91	-0.49	.008
Male	155	24.70	9.73			154	-17.53	17.29		
Race										
Black	43	26.14	9.70	0.00	.992	43	-21.38	14.24	-0.15	.361
White	114	26.12	9.59			113	-18.54	18.33		
MOS Type										
Com Support	45	24.89	10.01	0.33	.110	45	-21.36	17.95	-0.37	.072
Combat	49	21.57	9.93			49	-14.49	18.58		
Com Srv Sup	91	28.26	8.98	0.67	<.001	90	-19.74	15.57	-0.28	.078
Combat	49	21.57	9.93			49	-14.49	18.58		
Com Srv Sup	91	28.26	8.98	0.34	.049	90	-19.74	15.57	0.09	.589
Com Support	45	24.89	10.01			45	-21.36	17.95		
E5										
Gender										
Female	28	29.14	8.06	0.24	.222	28	-.59	13.14	-0.75	<.001
Male	174	26.81	9.54			175	10.38	14.62		
Race										
Black	50	27.76	9.57	0.10	.544	51	6.98	15.81	-0.26	.138
White	124	26.83	9.09			124	10.58	13.97		
MOS Type										
Com Support	54	25.27	9.60	-0.04	.849	55	10.76	15.94	-0.22	.343
Combat	46	25.63	9.09			46	13.40	11.92		
Com Srv Sup	99	28.78	9.15	0.35	.055	99	5.44	14.71	-0.67	.002
Combat	46	25.63	9.09			46	13.40	11.92		
Com Srv Sup	99	28.78	9.15	0.37	.027	99	5.44	14.71	-0.33	.039
Com Support	54	25.27	9.60			55	10.76	15.94		
Grade										
E6	97	30.94	8.78	0.41	.001	97	18.30	10.79	0.63	<.001
E5	209	27.17	9.28			209	8.94	14.78		
E6	97	30.94	8.78	0.53	<.001	97	18.30	10.79	2.13	<.001
E4	195	25.69	9.84			194	-18.78	17.44		
E5	209	27.17	9.28	0.15	.120	209	8.94	14.78	1.59	<.001
E4	195	25.69	9.84			194	-18.78	17.44		

Note. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* referent group. Referent groups (e.g., Whites) are listed second in each pair. Care should be exercised when interpreting the statistics because unequal cell sizes and interaction effects were not taken into account.

The pattern of results suggests the two scores are likely capturing relevant variance. Although there are score differences on the basis of gender, the pattern of differences is consistent with current differential MOS assignments within the Army that bar women from certain combat MOS. The General Experience score comprises experiences and activities more frequently encountered in combat MOS. Likewise, the Computer Experience score taps experiences more common to support MOS than combat MOS. Although the gender differences are not so large as to suggest bias, the differential opportunity for experiences across MOS suggests that use of the ExAct scores for promotion should be done on a within-MOS basis.

Revision of the ExAct for the Validation Data Collection

Given the large sample sizes needed (with both predictor and criterion data) to support the rainforest empiricism scoring method, it is unlikely that sufficient data will be collected in the validation effort to gain stable and reliable key values. Thus, the scoring scheme used for the field test analysis will be retained for the validation. Therefore, the ExAct was revised so soldiers would not have to distinguish between what transpired before and since their last promotion.

A review of the items suggested that the first 25 addressed activities and experiences for which relevance would likely fade over time. Therefore, the question posed to respondents on the validity data collection version of the instrument was changed from "How often have you performed each activity?" to "In the last 2 years, how often have you performed each activity?" No changes were made to individual items. These changes resulted in a single version of the ExAct suitable for all respondents in the validation effort. The version of the ExAct used for the validation effort is provided in Appendix C.

Operational Implementation Options and Issues

Using ExAct Scores in the Semi-Centralized Promotion System

There are multiple options for incorporating ExAct scores into the semi-centralized promotion system, each with associated strengths and weaknesses. Three options are presented below. This list is surely not exhaustive, but is suggestive of possible approaches.

The first option is to use the ExAct scores as reference material that commanders may consult to inform their evaluation. In the current promotion system, commanders assign up to 150 points towards each soldier's promotion point total. Commanders have a good bit of discretion in the assignment of these points, though in October 2000 more structure in the assignment of points was instituted.[7] In this approach, ExAct scores (along with guidelines or MOS norms) would be reported to commanders for their use in assigning points. The key advantage to this approach is that it would introduce the least amount of change into the promotion process. The key disadvantage is that the information would not be used in a standardized fashion or could be disregarded altogether.

[7] Specifically, commanders are now instructed to assign up to 30 points for each of the following areas: competence, military bearing, leadership, training, and responsibility and accountability.

Second, the ExAct scores could be used as reference material provided to Promotion Board members. This strategy is similar to the previous one in that the ExAct scale scores would be used only as reference information. The difference is that, rather than reporting scores to commanders, the information would be reported to the Promotion Board and used to inform their decisions. Again, the major advantage to this approach is that it would introduce little change to the current system. The information, however, would not be used in a standardized fashion and could be disregarded.

The approach with the most impact involves implementing the ExAct as a stand-alone instrument by which promotion points are directly assigned based on the scale scores. The ExAct scales would be scored and used to determine a set amount of promotion points. The key advantages to this approach are that use of the scores would be standardized and the scores would have direct impact on promotion decisions. The main disadvantage is that this approach would require a change to the current promotion point allocation system.

Possible Strategies for Completing an Operational ExAct

Regardless of the possible ways ExAct scores could potentially be used in the promotion system, someone will have to complete the form. Although the ExAct is a self-report form in the development and validation research, there are other options. Four possible options follow. Any of these options could conceivably work with either a written or automated form.

Administrative/archival records. In this option, the ExAct would be completed administratively through archival records. The key advantage to this approach is it would probably be immune to faking as the soldier would not be able to influence the results. This approach, however, would be difficult to implement because the ExAct requires information not currently available in Army records.

Supervisor. An alternative to the administrative approach is to have the soldier's supervisor complete the form. Again the key advantage is soldiers would be unable to directly engage in response distortion. However, several issues suggest this approach may not be feasible. First, supervisors would not be knowledgeable enough about their subordinates to complete the ExAct, given the detail of the items. Although supervisors presumably know what their subordinates' duties are, they may not be completely familiar with the soldiers' full history of relevant day-to-day and off-duty experiences. Also, the degree of familiarity is likely to vary—some supervisors have a small number of subordinates and may work very closely with them, whereas others head larger units and/or may not have an opportunity to work closely with their soldiers. To the degree a supervisor is not in a position to be highly familiar with a soldier's daily experiences, the amount of error in reporting could outstrip the potential problems associated with self-report.

Soldier self-report. The self-report method is being used in the development and validation research. Although this approach was deemed appropriate and useful for the research setting, the problem of response distortion becomes more salient in an operational setting. Although care was taken to make the items historical, external (i.e., observable behaviors), and verifiable in principle to reduce the potential for response distortion (Mael, 1991), the motivation to fake provided by the promotion context is likely to lead to some faking. Nonetheless, the

primary advantage of this approach is the soldier is best suited to complete the form. Further, allowing soldiers to control the information bearing on their promotion evaluation is likely to be more acceptable than if they had little or no control of the information.

Soldier and supervisor jointly. The last approach is designed to minimize the negative aspects of the other options while maximizing the positive aspects. In this approach, the ExAct would be completed jointly by the soldier and the supervisor. For instance, the soldier could provide the primary responses, and the supervisor would then review the form and "verify" its accuracy. Again, it is not expected supervisors would know everything about their soldiers' experiences, but they would be able to question suspect responses. For instance, suppose the soldier has indicated two years of experience as an instructor, but the supervisor doubts the soldier has ever been an instructor. The supervisor could question the soldier and ask for some form of verification. Similarly, a soldier may misunderstand some items or fail to realize some experience(s) as applicable and the supervisor may be able to help the soldier complete the form accurately. This approach would presumably be less work for the supervisors than if they were to complete the forms themselves but would provide a form of verification to minimize faking.

Although it is arguably the most reasonable approach of those described, a potential drawback is supervisors may not know about certain experiences and be unwilling to approve the form unless the soldier can validate his or her response. In addition to failing to receive due credit for their experiences, this may create conflict between soldiers and supervisors. In addition, there is no guarantee all supervisors would be equally sensitive to faked responses. Some supervisors may be more lenient when unsure whereas others may require proof for everything. This would be unfair because soldiers would not be treated equally. Despite these potential problems, however, the option of having the soldier and supervisor jointly complete the form is the most appealing of those discussed here.

Personnel File Form-21

Overview and Background

The design and content of the PFF21 are based largely upon the Project A Personnel File Form (PFF), the content of which was drawn primarily from the Army NCO Promotion Point Worksheet (PPW). Ordinarily, administrative personnel complete the PPW based on soldier records.

The decision to use a self-report instrument to gather relevant archival data for the current project, rather than collecting the information through administrative means, is based on the positive results of the Project A PFF development. In developing the PFF, it was found the archival records were not always current and often did not accurately reflect soldier information. Second, the self-report method provided the data substantially quicker and cheaper than was possible via administrative review of archival records. Third, the Project A development process was able to establish the necessary level of item specificity to collect accurate data from soldiers.

Instrument Development Process

The PFF21 was developed through a variety of means including adaptation of existing forms, workshops with NCOs, and pilot testing. To develop the initial framework, potential "information categories" were identified (e.g., civilian education, military training, awards), along with examples of specific pieces of information that would constitute each category. Several existing sources were used to develop the initial list, such as the NCO Promotion Point Worksheet, the Project A PFF, the NCO Evaluation Report (NCOER), and NCO Education System (NCOES) records and documentation.

Eighteen NCOs (E6-E9) were asked to evaluate the categories in terms of perceived relevance and acceptability for promotion decisions, and to add missing information. Overall, the workshop results indicated the information categories under consideration were acceptable and relevant. Although there was some objection to the civilian education category, including suggestions to remove it altogether, a majority of NCOs specifically stated no change was necessary. On the other hand, there was substantial concern that the proposed list of categories lacked information that would provide an assessment of MOS skills. Many NCOs voiced concern that the current system allows soldiers to be promoted who do not "know their job." Unfortunately, there are no operational indicators of MOS knowledge that could be added to the PFF21. Last, there was a perception of "unfairness" in the system. Through follow-up questions, however, it was established this criticism was directed at the larger promotion context in general and was not targeted towards the use of any particular type of information. That is, although soldiers perceive unfairness in the current promotion system, they generally perceived the information being used to make decisions as relevant and fair.

The prototype PFF21 developed from the workshops was pilot tested with 43 soldiers (n_{E6} = 41 and n_{E5} = 2) from Fort Riley. Soldiers were instructed to complete the form as accurately as possible and indicate any problems, confusing or vague items, or concerns either by writing on the form or talking to an administrator. No verbal or written comments were provided, indicating the form was clear and comprehended by the soldiers. Although responses could not be verified with archival data to check accuracy, an analysis of the responses showed almost all responses were within a reasonable, valid, range and consistent with the demography of the sample. A few soldiers reported a questionably high number of awards, memoranda/letters, and certificates.

Based on these results, the PFF21 appeared to be a suitable method of collecting archival information quickly and with reasonable accuracy. Minor changes were made to the form to enhance its use in the field test data collection. Given the success of collecting other archival data, items were added to collect self-report ASVAB re-test status (i.e., whether one has retaken the ASVAB) and ASVAB General Technical (GT) composite scores. Analyses of these scores are discussed in Chapter 6.

Scoring Considerations

Although the "rainforest empiricism" keying approach discussed previously could theoretically be used for the PFF21, the number of soldiers in the field test with criterion data (i.e., supervisor ratings) proved too small to gain stable and reliable key values. Moreover, we wanted to be able to simulate PPW scores as much as possible so that we could at least partially model the

current promotion system in the validation research. Therefore, scores on individual responses were weighted in a manner that corresponded as closely as possible to PPW specifications.

As with the ExAct, the PFF21 was administered to E6 soldiers to collect additional data that could be used as criteria if necessary. Again, to support this eventuality, E5 and E6 soldiers responded to items with regard to what occurred prior to their last promotion and what had occurred since their last promotion. Because it turned out this distinction was not required, E5 and E6 item responses were recalculated as necessary to simulate responses given with no time-based distinction. In some cases (e.g., with awards and college credits earned), this meant summing responses across time periods; in others, it meant taking the most recent response (e.g., weapons qualification).

Our interest in preserving comparability with the PPW meant we kept even those items that did not appear particularly useful. For example, many PFF21 items (e.g., specific awards earned) have low base rates. For such items, we computed only scale-level scores.

Scale Scores

Most scale scores were defined according to the way in which PPWs are scored. In cases where the Army differentially weights various accomplishments (e.g., specific awards, different types of military training), we attempted to compute scores weighted in the same manner. In some cases, we also report the unweighted scores. For PFF21 items that do not have an equivalent on the promotion worksheets, rational scales were identified and no item weighting system was imposed.

Thus, the scores listed below were derived from the PFF21. Scores that approximate PPW scores are indicated with an asterisk.

- Awards (weighted* and unweighted)
- Achievement Certificates
- PPW Achievement*
- Memoranda/Letters
- PPW Military Education*
- PPW Civilian Education*
- Disciplinary Actions
- Army Physical Fitness Test (APFT) (weighted and unweighted)
- Original PPW Weapons Qualification[8]
- Military Training*

Awards. This score was derived by summing the awards earned by each soldier. This score excluded awards not recognized on the PPW (e.g., the Physical Fitness Badge) and the "other" awards listed by the soldiers. Examination of these "other" awards indicated that they were not likely to be useful for making merit distinctions (e.g., the Overseas Ribbon that is given to any

[8] A recent change to the NCO promotion point worksheet has a more complicated method for obtaining this score that factors in, for example, the type of weapon used. We will use the simpler original formula because of limitations in what we can do with a self-report data collection format.

soldier who has an overseas tour). Two scores were generated – one using the Army's weighting system and the other assigning unit weights to each award. It is not clear how the Army's PPW weighting system was derived, but it assigns greater weight to the more prestigious awards.

Table 4.2 shows the descriptive statistics and subgroup analysis of the two Awards scores. As with the ExAct, subgroup analyses for the E6 sample were not conducted because the target population for the scores is E4 and E5 soldiers. The effect size for the E5 gender comparison is moderate, with women reporting fewer awards than did men. Both scores showed the expected increases with grade. The two award scores (weighted and unweighted) are highly correlated (r = .94). Because of its greater variance, the weighted score is likely to be a more useful predictor.

Achievement scores. The PPW gives credit for exceptional performance in training (e.g., Primary Leadership Development Course [PLDC] honor graduate) and for certificates of achievement. Using PPW weighting procedures, these indicators were combined to yield a "PPW Achievement" score. We also computed an Achievement Certificate score, that is the simple sum of certificates received, and a Memoranda/Letters score based on the number reported. Prior to computation of these scores, data were checked for outliers. This check showed only a few soldiers reporting a suspiciously high number of certificates or memos/letters. Responses greater than 15 were set to missing.

Table 4.3 shows the descriptive statistics and subgroup analyses for the Achievement Certificate and PPW Achievement scores. There were small gender differences, with males obtaining more certificates than females, and some differences in number of certificates earned based on MOS type. There was a gender difference on the PPW Achievement score for E5 soldiers, but not at the E4 level; the MOS-type differences were much less pronounced for this score than the Achievement Certificates score. Both scores showed the expected increases with grade. The Memoranda/Letters score (see Table 4.4) showed no subgroup differences except for the expected progression with increasing grade.

Military Education. A PPW Military Education score was computed by weighting the reported training courses (e.g., PLDC, Airborne School, Ranger School) according to PPW specifications. As shown in Table 4.5, E5 males scored higher than did females, and E4/E5 soldiers in combat service support MOS scored higher than those in combat MOS. Sample sizes were too small to examine gender or race differences at the E4 level. The Military Education score increased at each grade level.

Civilian Education. The hours reported for three types of school (i.e., college, trade, and business) were summed. Inspection of the distribution of responses showed a few soldiers reported more than 200 credit hours. A generous upper limit of 250 hours was imposed to clean the data of questionable responses. Consistent with the PPW, the number of credit hours was multiplied by 1.5 to derive the Civilian Education score. As shown in Table 4.6, there is a moderate gender difference in both the E4 and E5 samples. For both grades, women reported significantly more school credits than did men. In the E5 sample, blacks earned more credit than did whites. Among MOS types in both grades, soldiers in combat MOS reported less school credit than did soldiers in combat support and combat service support MOS. This difference could be confounded with gender differences. E6 soldiers had higher Civilian Education scores than did both the E5 and E4 soldiers; there was no difference between the E5 and E4 scores.

Table 4.2. Descriptive Statistics for the Unweighted and Weighted Awards

	Unweighted					Weighted				
Group	*n*	*M*	*SD*	Effect Size	*p*	*n*	*M*	*SD*	Effect Size	*p*
E4										
Gender										
Female	35	1.57	1.24	-0.29	.116	35	20.14	18.29	-0.24	.209
Male	162	1.99	1.45			162	24.38	18.01		
Race										
Black	45	1.71	1.38	-0.17	.332	45	22.33	19.76	-0.07	.691
White	120	1.96	1.48			120	23.63	18.04		
MOS Type										
Com Support	45	1.67	1.15	-0.24	.234	45	20.67	15.32	-0.28	.157
Combat	50	1.96	1.23			50	25.40	16.84		
Com Srv Sup	99	2.01	1.63	0.04	.849	99	24.24	20.27	-0.07	.729
Combat	50	1.96	1.23			50	25.40	16.84		
Com Srv Sup	99	2.01	1.63	0.30	.150	99	24.24	20.27	0.23	.294
Com Support	45	1.67	1.15			45	20.67	15.32		
E5										
Gender										
Female	29	2.90	1.05	-0.54	.005	29	40.52	17.85	-0.26	.195
Male	175	3.72	1.51			175	45.77	20.48		
Race										
Black	52	3.58	1.72	-0.04	.830	52	48.27	27.51	0.26	.321
White	124	3.63	1.35			124	44.19	15.96		
MOS Type										
Com Support	56	3.38	1.53	-0.08	.675	56	40.80	18.94	-0.16	.406
Combat	46	3.50	1.44			46	44.02	19.93		
Com Srv Sup	99	3.69	1.42	0.13	.464	99	47.17	20.62	0.16	.388
Combat	46	3.50	1.44			46	44.02	19.93		
Com Srv Sup	99	3.69	1.42	0.20	.204	99	47.17	20.62	0.34	.059
Com Support	56	3.38	1.53			56	40.80	18.94		
Grade										
E6	97	4.62	1.69	0.70	<.001	97	58.40	21.82	0.67	<.001
E5	210	3.60	1.46			210	45.02	20.00		
E6	97	4.62	1.69	1.90	<.001	97	58.40	21.82	1.91	<.001
E4	204	1.91	1.43			204	23.60	18.23		
E5	210	3.60	1.46	1.18	<.001	210	45.02	20.00	1.17	<.001
E4	204	1.91	1.43			204	23.60	18.23		

Note. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* referent group. Referent groups (e.g., Whites) are listed second in each pair. Care should be exercised when interpreting the statistics because unequal cell sizes and interaction effects were not taken into account.

Table 4.3. Descriptive Statistics for the Achievement Certificates and PPW Achievement

	Achievement Certificates					PPW Achievement				
Group	*n*	*M*	*SD*	Effect Size	*p*	*n*	*M*	*SD*	Effect Size	*p*
E4										
Gender										
Female	34	2.59	1.99	-0.37	.008	35	11.00	6.84	-0.10	.597
Male	146	3.77	3.21			162	11.76	7.85		
Race										
Black	40	3.30	2.63	-0.14	.446	45	11.22	8.13	-0.10	.588
White	110	3.74	3.24			120	11.96	7.62		
MOS Type										
Com Support	43	4.14	3.46	0.48	.088	45	13.44	6.64	0.24	.232
Combat	47	3.06	2.26			50	11.70	7.40		
Com Srv Sup	85	3.39	3.17	0.15	.497	99	10.20	8.11	-0.20	.275
Combat	47	3.06	2.26			50	11.70	7.40		
Com Srv Sup	85	3.39	3.17	-0.22	.222	99	10.20	8.11	-0.49	.020
Com Support	43	4.14	3.46			45	13.44	6.64		
E5										
Gender										
Female	28	3.71	3.38	-0.45	.027	29	14.31	8.21	-0.52	.012
Male	171	5.37	3.69			175	17.94	6.99		
Race										
Black	49	4.53	3.55	-0.24	.165	52	16.35	7.87	-0.25	.157
White	123	5.37	3.55			124	17.98	6.56		
MOS Type										
Com Support	52	5.87	3.87	-0.02	.906	56	17.56	7.26	-0.11	.591
Combat	45	5.96	3.62			46	18.37	7.31		
Com Srv Sup	99	4.36	3.46	-0.44	.013	99	16.72	7.39	-0.23	.211
Combat	45	5.96	3.62			46	18.37	7.31		
Com Srv Sup	99	4.36	3.46	-0.39	.016	99	16.72	7.39	-0.12	.479
Com Support	52	5.87	3.87			56	17.56	7.26		
Grade										
E6	95	6.63	4.20	0.40	.002	97	20.05	7.89	0.36	.004
E5	205	5.15	3.69			210	17.43	7.29		
E6	95	6.63	4.20	1.04	<.001	97	20.05	7.89	1.12	<.001
E4	185	3.48	3.04			204	11.35	7.79		
E5	205	5.15	3.69	0.55	<.001	210	17.43	7.29	0.78	<.001
E4	185	3.48	3.04			204	11.35	7.79		

Note. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* referent group. Referent groups (e.g., Whites) are listed second in each pair. Care should be exercised when interpreting the statistics because unequal cell sizes and interaction effects were not taken into account.

Disciplinary Actions. The Disciplinary Actions score was computed by summing the number of reported Article 15s and flag actions. Evaluation of raw frequencies showed a small number of extreme responses. These could be due to either false/careless responding or the

inclusion of outliers. To correct for these possibilities, responses indicating more than 15 Article 15s or flag actions were coded as missing. Results for the number of Disciplinary Actions are shown in Table 4.7. There is no evidence of subgroup differences on this score.

Table 4.4. Descriptive Statistics for the Memoranda/Letters Score

Group	*n*	*M*	*SD*	Effect Size	*p*
E4					
Gender					
Female	33	1.45	1.56	-0.30	.102
Male	143	2.33	2.96		
Race					
Black	37	1.81	1.90	-0.18	.196
White	108	2.37	3.08		
MOS Type					
Com Support	43	2.77	3.41	0.33	.208
Combat	46	1.98	2.40		
Com Srv Sup	83	1.96	2.56	-0.01	.975
Combat	46	1.98	2.40		
Com Srv Sup	83	1.96	2.56	-0.24	.139
Com Support	43	2.77	3.41		
E5					
Gender					
Female	27	3.37	2.94	-0.18	.379
Male	166	4.08	4.00		
Race					
Black	49	4.53	4.12	0.17	.340
White	118	3.90	3.81		
MOS Type					
Com Support	53	4.59	4.26	0.13	.530
Combat	44	4.05	4.11		
Com Srv Sup	93	3.61	3.34	-0.11	.513
Combat	44	4.05	4.11		
Com Srv Sup	93	3.61	3.34	-0.23	.129
Com Support	53	4.59	4.26		
Grade					
E6	96	5.96	4.37	0.50	<.001
E5	199	4.03	3.84		
E6	96	5.96	4.37	1.39	<.001
E4	181	2.15	2.75		
E5	199	4.03	3.84	0.68	<.001
E4	181	2.15	2.75		

Note. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* referent group. Referent groups (e.g., Whites) are listed second in each pair. Care should be exercised when interpreting the statistics because unequal cell sizes and interaction effects were not taken into account.

Table 4.5. Descriptive Statistics for the PPW Military Education Score

Group	*n*	*M*	*SD*	Effect Size	*p*
E4	107	7.44	7.68		
Gender					
Female	19				
Male	85	6.96	5.86		
Race					
Black	18				
White	69	6.55	5.27		
MOS Type					
Com Support	30	6.13	5.33	0.16	.616
Combat	29	5.52	3.92		
Com Srv Sup	44	9.45	10.28	1.00	.025
Combat	29	5.52	3.92		
Com Srv Sup	44	9.45	10.28	-0.62	.074
Com Support	30	6.13	5.33		
E5					
Gender					
Female	28	22.29	10.12	-0.33	.028
Male	172	27.37	15.55		
Race					
Black	50	27.44	16.34	0.07	.623
White	123	26.37	14.44		
MOS Type					
Com Support	53	25.89	10.16	0.18	.315
Combat	45	23.56	12.68		
Com Srv Sup	99	28.73	18.09	0.41	.051
Combat	45	23.56	12.68		
Com Srv Sup	99	28.73	18.09	0.28	.217
Com Support	53	25.89	10.16		
Grade					
E6	97	60.29	17.32	2.20	<.001
E5	206	26.72	15.29		
E6	97	60.29	17.32	6.88	<.001
E4	107	7.44	7.68		
E5	206	26.72	15.29	2.51	<.001
E4	107	7.44	7.68		

Note. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* referent group. Referent groups (e.g., Whites) are listed second in each pair. Care should be exercised when interpreting the statistics because unequal cell sizes and interaction effects were not taken into account.

Table 4.6. Descriptive Statistics for the PPW Civilian Education Score

Group	*n*	*M*	*SD*	Effect Size	*p*
E4					
Gender					
Female	30	90.70	76.58	0.64	.002
Male	130	41.60	76.86		
Race					
Black	33	38.64	72.77	-0.12	.539
White	101	48.19	78.81		
MOS Type					
Com Support	40	39.79	67.72	0.16	.478
Combat	42	29.36	64.69		
Com Srv Sup	78	62.48	81.14	0.51	.024
Combat	42	29.36	64.69		
Com Srv Sup	78	62.48	81.14	0.34	.132
Com Support	40	39.79	67.72		
E5					
Gender					
Female	27	75.78	64.78	0.50	.017
Male	163	44.30	62.69		
Race					
Black	47	65.68	66.82	0.48	.011
White	117	39.67	54.58		
MOS Type					
Com Support	50	52.59	66.16	0.73	.013
Combat	41	24.48	38.46		
Com Srv Sup	95	57.28	66.88	0.85	.001
Combat	41	24.48	38.46		
Com Srv Sup	95	57.28	66.88	0.07	.687
Com Support	50	52.59	66.16		
Grade					
E6	97	73.82	60.49	0.40	.001
E5	195	48.47	63.15		
E6	97	73.82	60.49	0.31	.009
E4	166	49.53	77.89		
E5	195	48.47	63.15	-0.01	.887
E4	166	49.53	77.89		

Note. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* referent group. Referent groups (e.g., Whites) are listed second in each pair. Care should be exercised when interpreting the statistics because unequal cell sizes and interaction effects were not taken into account.

Table 4.7. Descriptive Statistics for the Disciplinary Actions Score

Group	*n*	*M*	*SD*	Effect Size	*p*
E4					
Gender					
Female	34	0.53	.90	-0.20	.275
Male	150	0.78	1.26		
Race					
Black	40	0.80	1.36	0.15	.439
White	113	0.64	1.05		
MOS Type					
Com Support	45	0.69	1.18	-0.02	.939
Combat	48	0.71	1.24		
Com Srv Sup	88	0.67	1.09	-0.03	.854
Combat	48	0.71	1.24		
Com Srv Sup	88	0.67	1.09	-0.02	.929
Com Support	45	0.69	1.18		
E5					
Gender					
Female	28	0.29	0.76	-0.36	.061
Male	171	0.81	1.45		
Race					
Black	50	0.60	0.99	-0.15	.311
White	122	0.84	1.57		
MOS Type					
Com Support	53	0.75	1.24	-0.05	.805
Combat	44	0.82	1.28		
Com Srv Sup	98	0.87	1.82	0.04	.872
Combat	44	0.82	1.28		
Com Srv Sup	98	0.87	1.82	0.10	.688
Com Support	53	0.75	1.24		
Grade					
E6	96	0.76	1.22	-0.03	.766
E5	204	0.81	1.54		
E6	96	0.76	1.22	0.03	.794
E4	190	0.72	1.19		
E5	204	0.81	1.54	0.08	.506
E4	190	0.72	1.19		

Note. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* referent group. Referent groups (e.g., Whites) are listed second in each pair. Care should be exercised when interpreting the statistics because unequal cell sizes and interaction effects were not taken into account.

Army Physical Fitness Test (APFT) Score. The data were first cleaned for out-of-range responses, which were set to missing.[9] A weighted APFT score was computed from the conversion table used with the Army's PPW. Table 4.8 shows the descriptive statistics and subgroup analysis for the APFT test scores. As the table shows, most effect sizes are very small.[10] The exceptions are that soldiers in combat MOS outperform those in other MOS at the E4 level, and E5 soldiers tend to have somewhat higher scores than E4 soldiers. The weighted and unweighted APFT scores are correlated .84.

Weapons Qualification Score (WPN). Weapons qualification scores were computed according to the former Promotion Point Worksheet metric (Marksman = 10 points, Sharpshooter = 30 points, Expert = 50 points). Descriptive statistics and subgroup analyses are presented in Table 4.9. Males scored higher than females at both the E4 and E5 levels. It is possible that some of this difference is confounded with MOS type. As the table shows, soldiers in combat MOS scored higher than soldiers in combat support and combat service support MOS. Because women are not allowed in some combat MOS, the gender difference may be a result of differential opportunity. There is also some evidence that white soldiers outperform black soldiers on this measure. Soldiers in higher grades tend to have higher scores, with the exception of the E5/E6 comparison.

Military Training. The PPW calculates a Military Training score that is a composite of the weighted APFT and Weapons Qualification scores. This score was calculated using the available data, permitting a maximum of 100 points, as does the PPW. Table 4.10 shows the descriptive statistics for this score. The Military Training score shows a moderate effect size for gender (despite the different APFT score conversion tables for men and women) with both E4 and E5 females scoring lower, on average, than their male counterparts. Those soldiers in combat MOS (which generally exclude women) also outscore soldiers in both combat support and combat service support MOS. Soldiers in higher grades have higher scores, with the exception of the E5/E6 comparison.

Relationships Among Scores

Score intercorrelations, shown in Table 4.11, show a reasonable pattern of relationships. Closely allied scores are highly correlated (e.g., weighted and unweighted awards), but other scores have low to moderate correlations.

---

[9] Reported scores greater than 300 were erroneously set to missing in the field test analyses, but will be set to 300 in the validation analyses. Scores greater than 300 can be valid, but the PPW gives credit for no more than 300 points.

[10] To reflect physiological differences (e.g., aerobic capacity), some components of the APFT scores are based on different conversion tables for men and women and for different age groups. This likely minimized observed gender differences.

Table 4.8. Descriptive Statistics for the Unweighted and Weighted APFT Scores

	Unweighted					Weighted				
Group	*n*	*M*	*SD*	Effect Size	*p*	*n*	*M*	*SD*	Effect Size	*p*
E4										
Gender										
Female	33	244.85	54.34	-0.07	.714	33	25.09	12.37	-0.06	.739
Male	145	248.37	48.78			145	25.89	12.41		
Race										
Black	37	248.35	55.39	0.07	.736	37	26.43	12.76	0.12	.531
White	111	244.99	51.34			111	24.96	12.15		
MOS Type										
Com Support	43	241.81	51.09	-0.82	.009	43	23.72	12.44	-0.59	.008
Combat	48	264.71	28.01			48	30.52	11.55		
Com Srv Sup	84	243.31	56.24	-0.76	.015	84	24.79	12.08	-0.50	.009
Combat	48	264.71	28.01			48	30.52	11.55		
Com Srv Sup	84	243.31	56.24	0.03	.884	84	24.79	12.08	0.09	.643
Com Support	43	241.81	51.09			43	23.72	12.44		
E5										
Gender										
Female	28	260.57	30.91	0.06	.776	28	28.75	11.97	-0.04	.826
Male	167	257.76	50.62			167	29.25	10.98		
Race										
Black	48	257.27	55.48	-0.06	.785	48	29.46	10.57	0.06	.579
White	120	259.23	35.44			120	28.45	10.65		
MOS Type										
Com Support	50	254.66	55.38	-0.64	.102	50	28.52	10.93	-0.33	.166
Combat	43	269.67	23.55			43	31.44	8.93		
Com Srv Sup	97	254.59	52.58	-0.64	.074	97	28.43	11.89	-0.34	.140
Combat	43	269.67	23.55			43	31.44	8.93		
Com Srv Sup	97	254.59	52.58	-0.00	.994	97	28.43	11.89	-0.01	.966
Com Support	50	254.66	55.38			50	28.52	10.93		
Grade										
E6	93	256.98	40.86	-0.02	.849	93	28.15	11.02	-0.08	.841
E5	199	258.07	47.76			199	29.08	11.01		
E6	93	256.98	40.86	0.17	.162	93	28.15	11.02	0.17	.163
E4	184	248.67	49.24			184	26.03	12.36		
E5	199	258.07	47.76	0.19	.059	199	29.08	11.01	0.25	.011
E4	184	248.67	49.24			184	26.03	12.36		

Note. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* referent group. Referent groups (e.g., Whites) are listed second in each pair. Care should be exercised when interpreting the statistics because unequal cell sizes and interaction effects were not taken into account.

Table 4.9. Descriptive Statistics for the Weapons Qualification Score

Group	*n*	*M*	*SD*	Effect Size	*p*
E4					
Gender					
Female	33	22.12	14.95	-0.81	<.001
Male	145	34.28	16.02		
Race					
Black	39	28.46	15.48	-0.30	.100
White	110	33.45	16.45		
MOS Type					
Com Support	43	25.81	16.07	-1.24	<.001
Combat	48	42.50	13.45		
Com Srv Sup	84	24.76	12.08	-1.32	<.001
Combat	48	42.50	13.45		
Com Srv Sup	84	24.76	12.08	-0.07	.356
Com Support	43	25.81	16.07		
E5					
Gender					
Female	28	33.57	16.38	-0.93	.003
Male	172	43.84	11.10		
Race					
Black	50	39.60	14.14	-0.38	.039
White	123	43.82	11.20		
MOS Type					
Com Support	53	42.08	11.99	-0.31	.176
Combat	45	45.11	9.68		
Com Srv Sup	99	41.11	13.77	-0.41	.048
Combat	45	45.11	9.68		
Com Srv Sup	99	41.11	13.77	-0.08	.668
Com Support	53	42.08	11.99		
Grade					
E6	93	44.84	10.17	0.19	.115
E5	206	42.52	12.35		
E6	93	44.84	10.17	0.79	<.001
E4	184	31.74	16.50		
E5	206	42.52	12.35	0.65	<.001
E4	184	31.74	16.50		

Note. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* referent group. Referent groups (e.g., Whites) are listed second in each pair. Care should be exercised when interpreting the statistics because unequal cell sizes and interaction effects were not taken into account.

Table 4.10. Descriptive Statistics for the PPW Military Training Score

Group	*n*	*M*	*SD*	Effect Size	*p*
E4					
Gender					
Female	34	45.82	21.56	-0.51	.007
Male	151	57.77	23.37		
Race					
Black	41	50.93	22.44	-0.26	.147
White	113	57.09	23.44		
MOS Type					
Com Support	45	47.33	20.63	-1.28	<.001
Combat	48	73.02	20.05		
Com Srv Sup	88	50.59	21.79	-1.12	<.001
Combat	48	73.02	20.05		
Com Srv Sup	88	50.59	21.79	0.16	.408
Com Support	45	47.33	20.63		
E5					
Gender					
Female	28	62.32	18.62	-0.60	.004
Male	172	72.24	16.54		
Race					
Black	50	67.88	17.96	-0.25	.168
White	123	71.58	15.03		
MOS Type					
Com Support	53	68.98	16.19	-0.44	.049
Combat	45	75.16	14.10		
Com Srv Sup	99	68.97	18.72	-0.44	.050
Combat	45	75.16	14.10		
Com Srv Sup	99	68.97	18.72	0.00	.997
Com Support	53	68.98	16.19		
Grade					
E6	95	71.45	16.43	0.05	.688
E5	206	70.61	17.07		
E6	95	71.45	16.43	0.68	<.001
E4	191	55.65	23.37		
E5	206	70.61	17.07	0.64	<.001
E4	191	55.65	23.37		

Note. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* referent group. Referent groups (e.g., Whites) are listed second in each pair. Care should be exercised when interpreting the statistics because unequal cell sizes and interaction effects were not taken into account.

Table 4.11. PFF21 Score Intercorrelations

		1	2	3	4	5	6	7	8	9	10	11
1.	Awards											
2.	Weighted Awards	.94*										
3.	Achievement Certificates	.36*	.33*									
4.	PPW Achievement	.47*	.44*	.71*								
5.	Memos/Letters	.37*	.36*	.43*	.38*							
6.	PPW Military Education	.56*	.50*	.23*	.31*	.29*						
7.	PPW Civilian Education	-.00	-.00	.02	.06	.03	.16*					
8.	Disciplinary Actions	.01	-.01	.08	.04	.01	.02	-.03				
9.	APFT Score	.11*	.13*	.04	.13*	.05	.10*	.02	-.11*			
10.	Weighted APFT Score	.14*	.13*	.07	.13*	.05	.10	-.01	-.18*	.84*		
11.	Weapons Qualification	.39*	.36*	.22*	.27*	.29*	.23*	-.11*	-.00	.15*	.14*	
12.	Military Training Score	.36*	.33*	.19*	.29*	.23*	.23*	-.10*	-.09*	.57*	.69*	.80*

Note. Correlations based on E4 – E6 responses. Sample size ranges from 396 - 511.
* $p < .05$.

Preparation for the Validation Data Collection

A number of revisions, including the addition, deletion, and rewording of items, were made to the PFF21 to prepare it for the validation effort. Virtually all of these changes were made in an effort to approximate PPW scores more closely. As with the ExAct, the time distinction for E5 and E6 soldiers was dropped so that a single form will be suitable for all soldiers. The revised PFF21 is provided in Appendix D.[11] In addition to a more accurate derivation of the PPW-based scores obtained from the field test version, the revised PFF21 will allow computation of a Degree score (10 points for each academic degree earned) and a simulated PPW score (minus the Commander's Evaluation and Promotion Board appearance rating).

Implementing the PFF21 into the Semi-Centralized Promotion System

The possibility of implementing the PFF21 into the semi-centralized promotion system would not present significant difficulties. Promotions are handled within MOS, so the impact of gender score differences is minimized. Almost all of the information collected on the PFF21 is currently used in the promotion system in one form or another. As such, the PFF21 can best be viewed as an updated or revised component of the current PPW. The PFF21 form itself would not replace the PPW, but rather the PPW worksheet might be revised based on the information obtained from the PFF21 during the validation effort.

[11] Note that this version of the PFF21 also includes items for another ARI research project—an evaluation of the Army Continuing Education System (ACES).

CHAPTER 5: SEMI-STRUCTURED INTERVIEW

Background

The NCO21 semi-structured interview is designed to assess a soldier's standing on several important KSAs required for effective performance at the E5 and E6 grades. This interview package includes structured interview training and a standard protocol for conducting the interview, selecting questions from a question bank, developing new questions, and evaluating interviewees in several target areas.

The current method of selecting soldiers for promotion to the E5 and E6 NCO levels includes a semi-centralized board interview. Army regulations indicate this interview should cover the following six areas: Oral Communication, Military Presence/Bearing, Common Task Knowledge and Skill, Knowledge of World Affairs, Awareness of Military Programs, and Attitude. Two concerns about the board interview are (a) whether these six areas are the most important ones on which to focus and (b) a lack of standardization in the interview process because there is little formal guidance on how the board interview should be conducted. Indeed, the board interview is widely perceived as a formality rather than a requirement that screens soldiers for promotion. The NCO21 interview was designed to provide more structure to the process of identifying soldiers who are ready for advancement. It could conceivably be used to replace the board interview, but could also be used as an additional, separate component of the promotion process (e.g., as part of the NCO Educational System [NCOES]).

Instrument Design and Development Process

Project staff began designing the NCO21 interview by learning more about the traditional semi-centralized board interview. A focus group was conducted with 18 NCOs who had experience with these promotion boards. This meeting yielded pertinent information about the current promotion board appearance that guided the development of the semi-structured interview. For instance, some boards focus on job-specific (MOS) qualifications and others are less job-specific in their orientation. Overall, the focus group participants were concerned about soldiers getting promoted who do not "know their job" (i.e., MOS knowledge). Some participants also wanted promotion boards to focus on general leadership skills and traits—areas not explicitly measured in the current board interview.

Participants completed a brief survey with questions about semi-centralized promotion boards and their reactions to the idea of a more structured process. They indicated the optimal interview time to be 35 minutes; this is close to the estimated average time soldiers currently spend in board appearances (M = 28 min.; range = 8-45 min.). When asked about adverse effects they would expect if the individual appearance time required 45-60 minutes for each soldier, participants were divided (44% said it would not cause much of a problem, 17% did not know, and 39% believed it would cause a problem). Although the idea of implementing a semi-structured interview into the promotion process did not generate much enthusiasm, it was not rejected.

The initial stages of the semi-structured interview development process necessitated an examination of the extent to which the current promotion board interview assessed the expected future requirements of 21st-century NCOs (see Chapter 1). KSAs identified as not particularly

relevant to NCOs in the future would be excluded from the semi-structured interview for E4 and E5 soldiers. Instead, these areas would be replaced by a number of KSAs that target these expected requirements. High priority for inclusion in this measure was given to future-oriented KSAs not covered by other measurement methods and KSAs successfully measured by interviews in previous research. This initial screen ensured all KSAs applicable to 21st-century NCOs were covered in at least one instrument, and it allowed some KSAs to be measured by multiple methods, thus potentially minimizing method bias.

The current promotion board interview is intended to assess three KSAs expected to be applicable to 21st-century NCOs so these were initially included in the design for the new interview. Using the NCO21 nomenclature, these three KSAs are Oral Communication Skill, Military Presence, and Common Task Knowledge and Skill. Additional KSAs were given priority because they were not being assessed by any other predictor instruments (i.e., Adaptability and Problem-Solving/Decision Making) or because of their expected utility in an interview setting with an Army population (i.e., Level of Effort and Initiative on the Job; Level of Integrity and Discipline on the Job).

It was necessary to design the interview in a way that limited the number of KSAs and interview questions per KSA in order to provide the most meaningful information in a limited amount of time (i.e., 45-60 minutes). Interviews designed in previous studies (e.g., Peterson et al., 1999) typically allowed candidates 3-4 minutes to answer each question and asked two to three questions per KSA. Given this information, the NCO21 semi-structured interview was initially designed to measure six to seven KSAs. Later in the development process some KSAs were combined into KSA categories to expand the coverage of this assessment method.

The current promotion board interview asks questions primarily of a factual nature, in which there is a clear right and wrong answer. Although such knowledge-based questions can be useful, it is difficult to obtain a complete picture of how the soldier might perform on the job if the questions fail to measure the soldier's skills and aptitudes also. For this reason, the NCO21 semi-structured interview was designed to include three types of questions that collectively would tap important KSAs: (a) past-experience questions, (b) hypothetical situation questions, and (c) fact-based questions. Questions about past experiences ask the soldier to describe how he/she has behaved in certain types of situations; responses are intended to specify the situation that occurred, the soldier's action in response to the situation, and the result of the soldier's actions. Hypothetical situation questions present a fictitious but realistic scenario and ask the soldier to describe what he/she would likely do in that given situation. Similar to the promotion board questions, the fact-based questions are intended to have clear right answers, but the soldier is expected to describe how something should be done rather than provide a one-word answer. The intent was to maximize variability in the soldiers' responses.

Structured interviews developed in previous Army (Peterson et al., 1997) and government research projects were reviewed to identify materials that could be used to create a prototype interview (i.e., introductory script, questions, rating scales to evaluate a soldier's response). Existing questions were all of the "past-experience" type, and most contained probes or restatements of the question that could be used to elicit more details about the experience. Project staff generated four new past-experience questions. Collectively, 11 questions constituted the initial "question bank" from which an interviewer could draw. The questions assessed six

target KSAs: (a) Motivating, Leading, and Supporting Individual Subordinates; (b) Training Others; (c) Relating to and Supporting Peers; (d) Adaptability; (e) Level of Effort and Initiative on the Job; and (f) Level of Integrity and Discipline on the Job. Two additional KSAs—Oral Communication Skills and Military Presence—were also incorporated into the interview but would be evaluated based on observation only.

The rating scales from previous studies ranged from 1 (low) to 7 (high) and contained three anchor levels (i.e., low, moderate, and high), with short descriptions about general behavior demonstrated at each level. Each anchor level also included two to four specific behavioral examples of what the soldier could have described in his/her responses. There was no existing scale for Military Presence, so a draft scale was developed based on the KSA definition.

Other supporting materials developed for the prototype interview included an interview script, suggestions for probing, instructions for making ratings, and an interview worksheet to record ratings. The interview worksheet contained a list of the eight KSAs covered in the prototype interview, a place to record ratings (i.e., circle a value from 1 to 7), and a space to record notes.

Pilot Testing the Prototype Interview

Pilot testing of the prototype interview questions and rating scales included four sequential steps: (a) interview questions and rating scales from previous research were tested with E4 and E5 soldiers using a HumRRO interviewer, (b) senior NCOs wrote interview questions targeting 12 KSAs for potential inclusion in the question bank, (c) the expanded interview question bank and rating scales were pilot tested with E5 soldiers using a HumRRO interviewer, and (d) senior NCOs tried the interview process on each other. Each step is described below.

Project staff administered the draft semi-structured interview to eight E4 and E5 soldiers to evaluate the utility of the questions and accompanying rating scales. Each interview was limited to 15 minutes, so each interviewee responded to only a sample of the available questions. Each question was asked a total of three times. Responses to questions took less time than anticipated (i.e., 1-2 minutes per answer versus 3-4 minutes found in previous research). Some soldiers seemed reluctant to provide negative personal information regarding Level of Integrity and Discipline on the Job, so questions in this area were slightly more difficult to evaluate using the full range of the rating scale. The other questions yielded a variety of answers that could be evaluated reasonably well using the rating scales. None of the existing questions in the initial question bank were eliminated based on responses to the prototype interview. However, one question was set aside to be monitored for usefulness during the next data collection because it confused the soldiers.

In an effort to expand the question bank, a workshop was conducted with 16 senior NCOs (E7-E9) to develop new past-experience and hypothetical-situation interview questions. A secondary purpose of this workshop was to determine the feasibility of asking senior NCOs to develop new and usable questions for interviews, as this would be a part of the interview process. The participants were asked to write one past-experience question and one hypothetical-situation question each for 12 KSAs that potentially would be targeted in the interview. Half of the senior NCOs developed questions for the following KSAs: (a) Adaptability; (b) Team Leadership; (c) Problem-Solving/Decision Making; (d) Training Others; (e) Motivating, Leading, and Supporting Individual Subordinates; and (f) Directing, Monitoring, and Supervising Individual Subordinates.

The other participants were assigned the remaining six KSAs: (a) Cultural Tolerance, (b) Relating to and Supporting Peers, (c) General Self-Management Skill, (d) Self-Directed Learning Skill, (e) Level of Effort and Initiative on the Job, and (f) Level of Integrity and Discipline on the Job.

Overall, the senior NCOs were able to generate a few (i.e., two to seven) potentially usable questions for each KSA area. Project staff identified questions with potential based on question format, expected Army-wide applicability, and likelihood of potential variability in the responses. The senior NCOs were most successful at writing questions for Level of Integrity and Discipline on the Job and had less success writing questions for other KSAs (e.g., Problem-Solving/Decision Making). Project staff developed additional new questions for each of the KSAs based on critical incidents collected at the first focus group and from previous research efforts (e.g., Special Forces Project – see Russell, Crafts, Tagliareni, McCloy, & Barkley, 1996). In some cases, probes were developed for the new interview questions. Also, new rating scales were created for the six KSAs that were not assessed in the first pilot test (i.e., Team Leadership; Directing, Monitoring, and Supervising Individual Subordinates; Cultural Tolerance; General Self-Management Skill; Self-Directed Learning Skill; and Problem-Solving/Decision Making). These revisions increased the question bank from 11 questions covering 6 KSAs (average of 2 questions per KSA) to 58 questions covering 12 KSAs (average of 5 questions per KSA).

The revised interview package was pilot tested with 11 E5 soldiers using a HumRRO interviewer. Each interview lasted approximately 15 to 20 minutes, and each question in the question bank was asked at least twice. The HumRRO interviewer asked some of the soldiers to comment on the usefulness of the questions at the end of their interview. The interviewer also made assessments based on the types of responses obtained. These evaluations indicated four of the questions should be modified and five should be dropped from the question bank because they either confused the soldiers or yielded responses that lacked variance. The question bank and rating scales were revised based on the pilot test results and recommendations from project staff.

To assess the feasibility of using the semi-structured interview with a panel of Army interviewers, it was necessary to revise the interview package. First, the number of KSAs needed to be narrowed down to six or seven (plus Oral Communication Skill and Military Presence) so the interviewers could ask questions for each of the KSAs in the allotted time. Examinations of the question bank and responses to these questions from the earlier pilot tests determined that Problem-Solving/Decision Making and Cultural Tolerance had few useful questions and, in some cases, yielded responses that were difficult to rate using the rating scales. Because the SJT was expected to measure these constructs more effectively, they were dropped from the interview. Directing, Monitoring, and Supervising Individual Subordinates was also eliminated because it was measured particularly well by the SJT in the previous pilot tests. Some KSAs (e.g., Adaptability) were retained because they were not measured in any other predictor instrument. Common Task Knowledge and Skill and MOS/Occupation-Specific Knowledge and Skill were identified as important requirements for 21st-century NCOs, but interview time was limited. Therefore, only one of these areas (i.e., Common Task Knowledge and Skill) was included in the interview because it is assessed in the current semi-centralized promotion boards; as such, it would be useful to include this KSA should the interview replace the board interview.[12]

[12] Subsequently, the decision was made to replace Common Task Knowledge and Skill with MOS/Occupation-Specific Knowledge and Skill.

To maximize the number of KSAs covered, all remaining KSAs under consideration were included, but closely related KSAs were consolidated into categories. One KSA category was called Leadership Skills/Potential and included three leadership-related KSAs: (a) Motivating, Leading, and Supporting Individual Subordinates; (b) Training Others; and (c) Team Leadership. The second was labeled Self-Management and Self-Directed Learning Skill. Thus, the design of the panel interview included nine "target areas" (i.e., KSAs and KSA categories) seven of which had 38 total questions, plus Oral Communication Skill and Military Presence, which were assessed via observation during the course of the interview. New rating scales and definitions were developed for the KSA categories following the format of the previously tested scales. Table 5.1 summarizes the target areas that were pilot tested with HumRRO interviewers and the areas selected for the Army interviewer pilot test.

Table 5.1. Summary of KSAs Targeted During Pilot Testing

KSA	First Tryout (HumRRO Interviewer)	Second Tryout (HumRRO Interviewer)	Third Tryout (Army Interviewer)
Oral Communication Skills	✓	✓	✓
Military Presence	✓	✓	✓
Motivating, Leading, and Supporting Individual Subordinates	✓	✓	✓[a]
Training Others	✓	✓	✓[a]
Team Leadership		✓	✓[a]
Relating to and Supporting Peers	✓	✓	✓
Adaptability	✓	✓	✓
Level of Effort and Initiative on the Job	✓	✓	✓
Level of Integrity and Discipline on the Job	✓	✓	✓
Problem-Solving/Decision Making		✓	
Directing, Monitoring, and Supervising Individual Subordinates		✓	
Cultural Tolerance		✓	
General Self-Management Skill		✓	✓[b]
Self-Directed Learning Skill		✓	✓[b]
Common Task Knowledge and Skill			✓

[a]Consolidated into a KSA category titled Leadership Skills/Potential.
[b]Consolidated into a KSA category titled Self-Management and Self-Directed Learning Skill.

The panel interview training process and associated written training materials were designed so senior NCOs would have the opportunity to practice developing interview questions, conducting a simulated panel interview, evaluating the soldier, and consolidating the ratings. The training was not designed to be formal (i.e., extensive face-to-face training) but to provide a general structure for the process.

The Army tryout of the semi-structured panel interview was conducted with two groups of senior NCOs at USASMA (for a total of 20 E8-E9 NCOs). The senior NCOs familiarized themselves with the interview package and practiced selecting questions from the question bank and writing their own questions. Each participant was instructed to select one question from the question bank and write one new past-experience or hypothetical-situation question for each target area. They wrote two questions for Common Task Knowledge and Skill because there were no questions for this KSA in the question bank.

Participants formed interview panels with groups of four interviewers and one senior NCO role-playing the soldier being interviewed. The most senior NCO served as the lead interviewer. The lead interviewer appointed a recorder from the panel, who was responsible for consolidating the ratings at the end of the interview. All members of the panel were allowed to ask interview questions. In general, the "candidate" responses were curt; thus, each interview lasted an average of 20-25 minutes compared to the expected 45-60 minutes, with two to four questions asked per target area. Interviewers tended either to fail to take notes, or to take notes of an evaluative nature (e.g., "that was a good answer") rather than a more descriptive excerpt or quotation from the response. In terms of evaluating the responses, most participants did not appear to use the rating scale anchors. Further, none of the "recorders" on the panels computed the average ratings correctly. This suggested more instruction was required. In addition, nearly all of the individual ratings were 5 or higher on a 7-point scale, and there was low variance in ratings across target areas.

Following the interview exercise, the facilitator led a discussion to obtain feedback about the process. Many of the senior NCOs indicated the rating scales were too specific—describing trends or patterns of behavior difficult to measure in an interview setting. Instead, the participants suggested behavior-specific or situation-specific examples would be more effective. The discussion also revealed that appearing before the promotion board is a very formal and stressful experience and answers to questions tend to be terse. Even when interviewers probed (and they did not do so often), the responses were short and to the point. Interviewees did not provide detailed responses to reveal something particularly positive or negative about themselves. This is because, in traditional board appearances, soldiers are coached beforehand, and they are acutely aware they should only say things the board "wants to hear." Also, there is almost always someone on a board who knows and recommends the soldier for promotion; the other board members know this and are reluctant to score a soldier very low. Finally, several participants emphasized the importance of the soldier's record and demeanor during the interview over the soldier's responses to the interview questions. This is another reflection of the formal nature of traditional promotion board interviews.

In sum, the comments from the Army interviewer pilot test indicated that inherent in the semi-centralized promotion boards is a strong sense of tradition and culture, and an ingrained manner of thinking. Appearance before the board is viewed by most as a "rite of passage" for the NCO and is considered more of a formality for promotion than a means of selecting someone for advancement. If these characteristics of the traditional promotion board were to be carried over to the NCO21 semi-structured interview (as they were in this pilot test), the new interview would not likely be a significant improvement over the old in terms of reliably and validly differentiating among candidates for promotion. These factors were, therefore, considered when revising this measure prior to the field test.

Preparation for the Field Test

Results of the pilot test with Army interviewers clearly indicated a significant restructuring of the semi-structured interview package was warranted. In doing so, the key task was to shift the mindset associated with promotion boards to that of interviewing a candidate with the intent of obtaining specific, detailed responses illustrative of his/her knowledge, skills, and aptitudes.

Clearly, the senior NCO interviewers required more extensive training on the interview process. This was needed both to ensure the interviewers knew how to conduct the interview and use the rating scales *and* to emphasize they needed to adjust their mindset about how an interview should be conducted. Project staff designed the training to be more formal, using a face-to-face format to maximize learning and promote consistency in the manner in which interviews should be conducted. This format would allow the trainer an opportunity to emphasize the differences between the promotion board and the semi-structured interview. Based on previous interview research (e.g., Quartetti & Tsacoumis, 2000) and project team reviews, the training was revised to include the following components: (a) an explanation of the interview's relevance to future NCO promotions, (b) a description of the target areas, (c) instructions on how to prepare the interview, (d) instructions on how to use prepared questions and develop new questions, (e) instructions on how to design the interview, (f) guidance on conducting the interview, (g) tips for asking questions, (h) instructions on how to evaluate the soldier, (i) a practice exercise on evaluating candidates, and (j) a practice exercise on preparing for and conducting an entire interview. Both practice exercises incorporated a feedback component so common errors in writing new questions, evaluating candidates, and consolidating ratings could be addressed and corrected prior to administering interviews in the field test.

Multiple interviewers were considered necessary to help assure reliable and accurate measurement. Practical constraints, however, limited consideration of more than two interviewers per soldier. Therefore, the field test used two-person interviewer teams.

Another challenge was to combat the interviewers' propensity to render ratings that were too lenient (and limited in scope). To do so, the rating scales were revised so they were more targeted – eliminating the general paragraph descriptions describing typical behavior patterns and focusing on more concrete behaviors that might be evident in a candidate's response. The revised rating scales (used both in the field test and validation data collection) are shown in Appendix E. Training materials were also revised to provide more detail about how to take effective notes and use the rating scales to evaluate the candidate.

Next, Common Task Knowledge and Skill was replaced with MOS/Occupation-Specific Knowledge and Skill because project staff learned the Army currently uses local exams to assess common task knowledge and skill. Recall from the initial focus group that participants were concerned about a lack of proficiency in the soldier's MOS, and some wanted to assess this construct in the board interview. A rating scale was created for the MOS/Occupation-specific KSA. Similar to Common Task Knowledge and Skill questions, interviewers would be required to develop their own MOS-specific interview questions to ask of candidates in the same MOS. Field test interviewers were instructed not to ask MOS-specific questions of soldiers who were not in their MOS.

The question bank was expanded to include several questions from the USASMA pilot test, and the project team, for a total of 53 questions. Table 5.2 shows that nearly half (45%) of the revised question bank included past-experience questions, and 55% were hypothetical-situation questions.[13] Note that hypothetical-situation questions were found to be more conducive to assessing interview performance in some categories (e.g., Adaptability, Leadership Skills/Potential) than others (e.g., Self-Management/Self-Directed Learning, Relating to Peers). The question bank was also modified to more clearly separate the questions appropriate only for E5 soldiers.

Table 5.2. Composition of Field Test Interview Question Bank

Target Area	Total Number of Questions in Bank	Number of Past-Experience Questions	Number of Hypothetical-Situation Questions
Adaptability	9	2	7
Self-Management/Self-Directed Learning	5	5	0
Level of Effort & Initiative on the Job	5	2	3
Level of Integrity & Discipline on the Job	10	3	7
Relating to & Supporting Peers	7	5	2
Leadership Skills/Potential	17	7	10
Total	53	24	29

Note. The field test structured interview also assessed MOS/Occupation-Specific Knowledge and Skill, Oral Communication Skills, and Military Presence. The latter two are evaluated by observation and MOS/Occupation-Specific Knowledge and Skill questions are generated by the interviewers.

Finally, given the difficulties in recording and consolidating ratings during the USASMA pilot test, the summary worksheets were revised to facilitate an accurate assessment of the candidate's interview performance based on ratings from two interviewers.

Field Test Administration

The revised semi-structured interview and training materials were field tested at three Army installations. At each site, two project staff members facilitated a 3.5- to 4-hour face-to-face training session with senior NCOs who would serve as interviewers. Training participants included 10 senior NCOs at Fort Carson, 6 at Fort Stewart, and 8 at Fort Leonard Wood.

Across the three sites, 68% of the interviewers were white, 92% were male, and a small majority (45%) were in combat support MOS. Each participant received training materials required to understand and conduct the interview: (a) definitions of target areas, (b) question bank, (c) question development worksheet, (d) interview introduction script, (e) individual rating worksheet, (f) rating scales, and (g) a worksheet to record and consolidate consensus ratings from both interviewers.

The senior NCOs participated in two practice exercises to familiarize themselves with the interview process. Following each exercise, staff trainers led a discussion to review participants'

[13] Fact-based questions are not suitable for the question bank because in an operational setting this would easily result in compromise.

performance on the exercise and provide feedback to minimize common errors that occurred during the pilot tests (e.g., errors in making ratings). Toward the end of the training session, the training facilitator assigned the interviewers to pairs. No interviewer pair members were of the same MOS because of the variety of MOS across interviewers. Each pair remained interview partners for the duration of the data collection period. The most senior NCO served as the lead interviewer, responsible for making introductions, explaining the process to the candidate, and making the final decision on selecting interview questions. The second interviewer was designated the recorder, responsible for consolidating the ratings at the end of the interview. Both interviewers could ask questions and were instructed to take notes during the interview. At the end of the interview, both interviewers were required to review their notes and make independent judgments using the target area rating scales. If their ratings differed by more than 2 points the interviewers discussed the candidate's performance and reached consensus within 2 points. The two sets of ratings were then averaged to obtain an overall rating for each target area. An overall rating was computed by averaging the final target area ratings.

All interviews were conducted with E4 and E5 soldiers during their respective written testing sessions. One staff member of the project team served as Interview Manager. This individual designated which soldiers should be interviewed, when possible, based on a match between the MOS of an interviewer and the soldier (thus allowing the interviewer(s) to ask MOS-specific questions). Each interview lasted approximately 20 minutes, with 10 extra minutes built into the process for completing the rating forms. After completing their interviews, soldiers returned to the written testing session. After all interviews were conducted, the senior NCO interviewers were asked to complete a form to evaluate and provide feedback on the interview and training.

Before the results of the field test are presented, some caveats should be noted. One interviewer did not return on the second and third day to interview soldiers. Consequently, the HumRRO staff person who led the interview training filled in for the interviewer and served as the recorder for 3 out of 4 sessions. Finally, some interviewers rated soldiers on MOS-specific knowledge and skill even though they were not of the same MOS. Such ratings can be valid if the MOS of the interviewer and soldier were similar or if the question was general enough to apply to more than one MOS. The extent to which invalid ratings in this area occurred is unclear. Therefore, all MOS-specific ratings were retained for the field test analyses.

Field Test Results

The field test data yielded four sets of scores on a 7-point scale for each E4 or E5 soldier participating in the interview: one score for each of the nine target areas from each interviewer, a set of average consensus ratings (i.e., agreement within 2 points) for each target area, and an overall interview score (i.e., composite of the average consensus ratings). The following section presents the results from the descriptive statistics, analysis of subgroup differences, scale building analyses, reliability estimates, and participant feedback.

Descriptive Statistics

A total of 210 interviews were completed during the field test. Of these soldiers, 103 (49%) were E4s and 105 (51%) were E5s (grades for two interviewees are unknown). Table 5.3 shows the averaged consensus ratings for each target area as well as the overall average

interview scores (i.e., composite scores). A composite score excluding the MOS-specific rating was also computed, as most soldiers were not rated in this area and soldiers rated in this area were primarily evaluated by only one interviewer. Overall, the amount of variability in the ratings suggests interviewers were able to discriminate among soldiers. Although the mean values (4.8-5.3) indicate some evidence of leniency in the ratings, these means are much lower than those from the Army interviewer pilot test.

Table 5.3. Descriptive Statistics for Interview

Target Area	*M*	*SD*
Adaptability	5.21	0.99
Self-Management and Self-Directed Learning Skill	4.94	1.12
Level of Effort and Initiative on the Job	5.25	0.94
Level of Integrity and Discipline on the Job	5.16	1.19
Relating to and Supporting Peers	5.03	0.98
Leadership Skills/Potential	4.99	0.98
MOS/Occupation-Specific Knowledge and Skill	4.77	1.38
Oral Communication Skill	5.19	1.04
Military Presence	5.22	1.06
Composite Interview Score	5.11	0.81
Composite Interview Score Excluding MOS-Specific Ratings	5.12	0.80

Note. $n = 210$ for all variables except Leadership Skills/Potential ($n = 209$) and MOS/Occupation-Specific Knowledge and Skill ($n = 83$). Interviewers' average consensus ratings ranged from 1.5-7.0 for the target areas.

Analysis of Subgroup Differences

To maximize sample size, subgroup analyses used the composite interview score without the MOS-specific rating. Table 5.4 shows the subgroup differences by soldier gender, race, grade, and MOS type. The analyses revealed no significant differences in interview scores by gender, race, or MOS type. As expected, E5 soldiers received significantly higher composite interview scores than did E4 soldiers.

Scale Building

Average consensus scores for each target area were correlated to assess the relationships among the scores. Table 5.5 shows that all scores were significantly intercorrelated, $p < .001$, with the exception of the lack of relationship between MOS/Occupation-Specific Knowledge and Skill and Level of Integrity and Discipline on the Job, $r = .14$, $p = .21$. The high correlations suggested that the semi-structured interview measured a single underlying construct.

An exploratory factor analysis (EFA) was performed to determine if the semi-structured interview assessed more than one underlying construct. The EFA used a maximum likelihood extraction with oblique rotation. Models with two, three, and four factors were tested, both including and excluding the MOS rating. The analyses did not reveal meaningful results (i.e., no simple structure and high correlations among factors). These results, coupled with the high inter-

item correlations, suggested the semi-structured interview measures one underlying construct. We concluded that the overall composite score is the most appropriate summary score for the interview.

Table 5.4. Subgroup Differences in Composite Interview Scores

Group	*n*	*M*	*SD*	Effect Size	*p*
Gender					
Female	33	5.28	0.81	0.23	.228
Male	170	5.09	0.81		
Race					
Black	46	5.20	0.77	0.14	.396
White	130	5.08	0.86		
Pay Grade					
E5	105	5.25	0.72	0.28	.036
E4	103	5.01	0.87		
MOS Type					
Combat Support	49	4.94	0.87	-0.22	.308
Combat	45	5.11	0.77		
Combat Service Support	103	5.22	0.80	0.14	.451
Combat	45	5.11	0.77		
Combat Service Support	103	5.22	0.80	0.32	.051
Combat Support	49	4.94	0.87		

Note. Subgroup differences in composite interview score excluding MOS/Occupation-Specific Knowledge and Skill ratings. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* referent group. Referent groups (e.g., Whites) are listed second in each pair. Care should be exercised when interpreting the statistics because unequal cell sizes and interaction effects were not taken into account.

Table 5.5. Inter-Item Correlations for the Semi-Structured Interview (Consensus Ratings)

Rating	1	2	3	4	5	6	7	8	9
1. Adaptability	--								
2. Self-Management and Self-Directed Learning	.56	--							
3. Level of Effort and Initiative on the Job	.59	.60	--						
4. Level of Integrity and Discipline on the Job	.45	.44	.57	--					
5. Relating to and Supporting Peers	.53	.48	.58	.51	--				
6. Leadership Skills/Potential	.48	.46	.62	.47	.62	--			
7. MOS/Occupation-Specific Knowledge and Skill	.43	.44	.54	.14	.47	.65	--		
8. Oral Communication Skill	.58	.48	.63	.43	.63	.69	.63	--	
9. Military Presence	.58	.47	.66	.44	.53	.63	.51	.73	--

Note. n = 209-210 except for correlations with MOS/Occupation-Specific Knowledge and Skill score (n = 83). All correlations, except .14, are significant at $p < .001$.

Reliability Estimates

Internal consistency reliability (Cronbach's alpha) was computed for the two composite scores. Computing the composite without the MOS-specific rating offered a greater sample size for the computation of internal consistency reliability estimates. Alpha was .92 (n = 82) for the composite score including all target area ratings and .91 (n = 209) for the composite score excluding MOS-specific ratings. Based on this analysis, there was no evidence to suggest that any target areas should be dropped from the semi-structured interview.

The first step in estimating interrater reliabilities was assigning one member of each interviewer pair to one rater group and the other member to a second rater group. Reliabilities were estimated by computing the (a) correlation between the two groups of ratings for each target area and composite (r) and (b) intraclass correlation coefficients (*ICCs*; Shrout & Fleiss, 1979, *ICC*[3,1] assessing consistency across a fixed set of raters). These reliabilities were computed for the ratings both before and after consensus. Table 5.6 shows the interview pairs tended to provide consistent (i.e., reliable) ratings for each target area and composite score before and after consensus. Consensus ratings were only slightly more reliable than individual ratings rendered prior to consensus discussions.

Table 5.6. Interview Interrater Reliability Estimates

Target Area	Interrater Reliability (single rater)			
	Before Consensus		After Consensus	
	r	*ICC*	*r*	*ICC*
Adaptability	.57	.56	.62	.61
Self-Management and Self-Directed Learning	.65	.65	.65	.65
Level of Effort and Initiative on the Job	.62	.62	.62	.62
Level of Integrity and Discipline on the Job	.65	.65	.68	.68
Relating to and Supporting Peers	.59	.59	.62	.61
Leadership Skills/Potential	.59	.59	.61	.60
MOS/Occupation-Specific Knowledge and Skill	--	--	--	--
Oral Communication Skill	.64	.64	.67	.66
Military Presence	.62	.62	.65	.65
Composite Interview Score	.78	.78	.80	.80
Composite Interview Score Excluding MOS-Specific Ratings	.77	.77	.79	.79

Note. Sample sizes range from 207-209. One member of each interviewer pair was assigned to one rater group and the other member to a second rater group; reliabilities were estimated by computing the (a) correlation between the two groups of ratings for each target area and composite (r) and (b) intraclass correlation coefficients (*ICCs*; Shrout & Fleiss, 1979, *ICC*[3,1] assessing consistency across a fixed set of raters). MOS/Occupation-Specific Knowledge and Skill not included in this analysis because generally only one interviewer matched the interviewee on MOS and made this rating.

Summary of Participant Feedback

After all interviews were conducted at each field test site, the senior NCO interviewers were asked to evaluate the semi-structured interview. Participants indicated the extent to which they were satisfied with the various components of the interview using a 5-point scale ("not at all" to "a very great extent"). The data suggested the interviewers were generally satisfied with the interview and considered it to be at least moderately useful to the E5/E6 promotion process (Table 5.7). The data suggested no major problems with the interview or the training. The interviewers were also encouraged to provide written feedback about the interview. Written comments were few, but they primarily addressed specific questions in the question bank.

Table 5.7. Interview Evaluation Results

Components of the Interview	Percent Responding		
	Not at All/Slight Extent	Moderate Extent	Great Extent/Very Great Extent
1. This structured interview would provide useful information to the E5/E6 promotion process.	13.6	54.5	31.8
2. The training was sufficient preparation for conducting these interviews.	4.5	31.8	63.6
3. The definitions of the *Performance Areas* are clear and concise.	8.7	13.0	78.2
4. The soldiers/interviewees understood the questions that were selected from the *Question Bank.*	13.0	26.1	60.8
5. The soldiers/interviewees understood the questions that my interview pair developed.	8.7	21.7	69.6
6. Writing new questions was manageable.	4.3	30.4	65.2
7. The rating scale anchors were useful for evaluating interviewee responses to questions.	4.5	40.9	54.5
8. The Overall Average Score on the *Interview Summary Worksheets* accurately reflected my overall evaluation of the candidates' structured interview performances.	0.0	34.8	65.2

Note. n = 22-23.

Preparation for the Validation Data Collection

Several steps were taken to revise the materials and procedures for conducting the interviews in the validation study. Based on the interviewers' comments, some questions were dropped from the bank because they were too easy and other questions were edited slightly either to clarify or make them applicable for both E4 and E5 soldiers. In addition, five questions developed by the Army interviewers during the field test were added to the question bank. Table 5.8 summarizes the content of the validation data collection structured interviews, listing the final target areas and the number of questions per area that are included in the question bank. In addition, scannable forms were developed for two interview forms to facilitate data entry: (a) the worksheet for recording notes and individual ratings, and (b) the interview evaluation form.

The method of conducting interviews was also altered to maximize the number of interviews that could be collected in the validation study. Rather than removing soldiers from the paper-and-pencil sessions to participate in the interview, soldiers will be scheduled for interviews in a separate session. Approximately five pairs of interviewers will conduct interviews in both morning and afternoon sessions on each day of the data collection. Interviews will be scheduled for 45 minutes.

Table 5.8. Summary of Validation Data Collection Interview Target Areas and Questions

Target Area	Total Number of Questions in Bank	Number of Past Experience Questions	Number of Hypothetical Situation Questions
Adaptability	9	2	7
Self-Management/Self-Directed Learning	6	5	1
Level of Effort & Initiative on the Job	4	2	2
Level of Integrity & Discipline on the Job	11	3	8
Relating to and Supporting Peers	7	5	2
Leadership Skills/ Potential	13	6	7
MOS-Specific Knowledge and Skill	Interviewer Writes	Interviewer Writes	Interviewer Writes
Oral Communication Skill	N/A	--	--
Military Presence	N/A	--	--
Total	50	23	27

Note. Some target area labels are abbreviated.

CHAPTER 6: OPERATIONAL PREDICTOR MEASURES

Three measures used as operational predictors in other contexts were selected for use as experimental predictors in the NCO21 research effort. The Armed Services Vocational Aptitude Battery (ASVAB) is used to help select and classify enlisted personnel upon initial entry into the U.S. military service. The Assessment of Individual Motivation (AIM) is being used as a supplemental screen in a pilot test for enlisted applicants who do not have a high school diploma. Finally, the Biographical Information Questionnaire (BIQ) comprises items from several existing biodata instruments that are used for various purposes (e.g., screening soldiers interested in joining the Special Forces).

Armed Services Vocational Aptitude Battery (ASVAB)

Description and Operational Uses

The ASVAB has 10 subtests (see Table 6.1). Composites derived from various combinations of these subtests are used to make entry-level selection and classification decisions for enlisted personnel. With regard to selection, all U.S. military services use the Armed Forces Qualification Test (AFQT), which is derived from the Word Knowledge, Paragraph Comprehension, Arithmetic Reasoning, and Math Knowledge subtests. Qualification for various MOS is based on Aptitude Area scores that are a set of Army-specific composite scores computed from ASVAB subtests and designed to predict success in various types of MOS.

Table 6.1. ASVAB Subtests

Arithmetic Reasoning
Numerical Operations
Paragraph Comprehension
Word Knowledge
Coding Speed
General Science
Mathematics Knowledge
Electronics Information
Mechanical Comprehension
Automotive-Shop Information

Certain ASVAB scores are also relevant for in-service personnel decisions. In particular, the General Technical (GT) composite score determines eligibility for reenlistment options, MOS changes, and for a number of advanced MOS designations (e.g., Special Forces MOS). The GT score is based on three of the four AFQT subtests (Word Knowledge, Paragraph Comprehension, and Arithmetic Reasoning). Because the GT score is used in-service, soldiers are likely to know this score without having to look it up. Further, soldiers with relatively low GT scores based on their pre-enlistment ASVAB will often retake the ASVAB in an effort to increase their GT scores. When the ASVAB is retaken in-service, it is referred to as the Armed Forces Classification Test (AFCT). For ease of discussion, we will use the term "ASVAB" throughout this report to refer to both pre-enlistment and in-service administrations of this test battery.

NCO21 Project Application

The ASVAB assesses several relevant NCO21 KSAs, particularly General Cognitive Aptitude, which in turn should be related to some other KSAs identified as critical to measure in Phase II of this project (e.g., Problem-Solving/Decision Making, Knowledge of the Inter-Relatedness of Units). Two alternative composite scores will be used to indicate General Cognitive Aptitude – the AFQT and the GT composites. By virtue of the fact that GT scores contribute to the determination of reenlistment options, the current NCO promotion system indirectly factors GT score into the promotion decision. AFQT and GT scores of record will be retrieved from Army records. Additionally, the PFF21 asks soldiers to report their latest GT score and to indicate how often they have retaken the ASVAB. In the NCO21 field test, 31% of the soldiers indicated they had retaken the ASVAB. The majority of these soldiers (86%) had retaken the ASVAB only once. Correlations between the AFQT score of record and the soldier's most recent GT score (self-reported) in the field test sample are shown in Table 6.2.

Table 6.2. AFQT and Self-Report General Technical Score Intercorrelations

Paygrade	*n*	*r*
E4	179	.39
E5	196	.56
E6	90	.64

Note. All correlations significant at $p < .01$.

The question of how often a soldier has retaken the ASVAB in an effort to improve his/her performance is a potentially important one. The Army allows soldiers to take the test four times – once at pre-enlistment with a maximum of three retests. Generally, soldiers retest after taking steps, such as education and training, to improve their scores. It is conceivable that the predictive value of scores based on different administrations will be different. Specifically, we would expect the predictive validity of the first score earned would be higher than the validity of a score based on repeated attempts. Therefore, we would be interested in comparing ASVAB scores obtained in any retakes with the pre-enlistment scores.

A complication in making the desired comparisons is obtaining the required data. Army records do not consistently provide accurate information on the number of retakes and normally write over previous scores when retake scores are recorded. Therefore, we obtained ASVAB-related data using the following strategies:

- Retrieved pre-enlistment ASVAB scores from Army accession files.
- Retrieved all available ASVAB score information from the Army's enlisted master files (EMF).[14]
- On the PFF21, asked soldiers to report the number of times they have retaken the ASVAB.
- On the PFF21, asked soldiers to report their current GT score of record.

[14] In the validation effort, other background information, such as soldier time in service, gender, and race/ethnic group will also be retrieved from or computed based on information in the enlisted master files.

In the field test analyses reported in Chapter 7, the AFQT scores were retrieved from the Army enlisted master files and the GT scores were self-reported. In the validation data collection, there will be additional effort to identify AFQT and GT scores that are based on pre-enlistment testing versus one or more retests. We can then contrast the predictive validity of the pre-enlistment and retest scores.

Finally, it is relevant to note that Army applicants who take the Computerized Adaptive Testing version of ASVAB (CAT-ASVAB) receive an 11th subtest called Assembling Objects. This test is currently being administered on an experimental basis and scores for NCO21 research participants are not likely to be available. For future reference, however, Assembling Objects could be used to assess two other NCO21 KSAs – Spatial Relations Aptitude, and Perceptual Speed and Accuracy.

Assessment of Individual Motivation (AIM)

Description and Operational Use

AIM reliably measures six temperament constructs: Dependability, Adjustment, Work Orientation, Leadership, Agreeableness, and Physical Conditioning (White & Young, 1998; Young, Heggestad, Rumsey, & White, 2000). Several of these constructs are similar to some of the NCO21 KSAs. In the Army's Project A research these temperaments were measured by a self-report instrument called the Assessment of Background and Life Experiences (ABLE). The Project A results, involving nearly 60,000 enlisted personnel, established that individual differences in soldiers' Work Orientation, Leadership, and Dependability as measured by ABLE are important determinants of the duty performance of NCO and first-term enlisted personnel (J. Campbell & Knapp, 2001; Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; Rumsey, Peterson, Oppler, & J. Campbell, 1996; White, Young, & Rumsey, 2001). Soldiers' levels of Adjustment, Physical Conditioning, and Dependability (as measured by ABLE) were also predictive of attrition during the first term of enlistment (White, Nord, Mael, & Young, 1993). The ABLE's scales showed little overlap with ASVAB, either conceptually or statistically.

There was much interest in using ABLE for enlisted personnel selection and classification decisions but its proposed implementation was withdrawn largely due to concerns about its susceptibility to deliberate faking (White et al., 2001). Deliberate faking leads to elevated test scores that are of little value for personnel decisions. Given these concerns, AIM was developed by ARI to measure the performance-relevant constructs from ABLE with greater resistance to deliberate faking. Definitions for the AIM scales are shown in Table 6.3.

Each AIM item consists of four behavioral statements (tetrad) that are indicative of the underlying psychological constructs being measured. For each item examinees are asked to identify which statement is most and least descriptive of them. To reduce AIM's susceptibility to deliberate faking, the self-statements within the tetrad are balanced in terms of social desirability. The AIM contains 27 items and takes about 25 minutes to complete. Seven items were added to AIM for the NCO21 field test to augment its internal consistency reliability.

Table 6.3. Definitions of the AIM Scales

Title	Definition
Work Orientation	The tendency to strive for excellence in the completion of work-related tasks. Persons high on this construct seek challenging work activities and set high standards for themselves. They consistently work hard to meet these high standards.
Adjustment	The tendency to have a uniformly positive affect. Persons high on this construct maintain a positive outlook on life, are free of excessive fears and worries, and have a feeling of self-control. They maintain their positive affect and self-control even when faced with stressful circumstances.
Agreeableness	The tendency to interact with others in a pleasant manner. Persons high on this construct get along and work well with others. They show kindness, while avoiding arguments and negative emotional outbursts directed at others.
Dependability	The tendency to respect and obey rules, regulations, and authority figures. Persons high on this construct are more likely to stay out of trouble in the workplace and avoid getting into difficulties with law enforcement officials.
Leadership	The tendency to seek out and enjoy being in leadership positions. Persons high on this scale are confident of their abilities and gravitate towards leadership roles in groups. They feel comfortable directing the activities of other people and are looked to for direction when group decisions have to be made.
Physical Conditioning	The tendency to seek out and participate in physically demanding activities. Persons high on this construct routinely participate in vigorous sports or exercise, and enjoy hard physical work.

The AIM has been shown in a series of investigations to predict measures of soldiers' duty performance and adaptability with comparable or higher criterion-related validity than the ABLE. In addition, preliminary findings indicate that AIM is more resistant to deliberate faking than ABLE (Young et al., 2000; White & Young, 2001). Several AIM scales have been found to be predictive of soldiers' attrition during their first term of enlistment (Young et al., 2000). In other research, Work Orientation and Leadership were linked to Special Forces field performance with validity estimates of .21 to .30 (Kilcullen, Chen, Zazanis, Carpenter, & Goodwin, 1999). Work Orientation and Dependability were also strongly associated ($R = .44$) with the successful performance of Correctional Specialists (MOS 95C) who guard inmates in the DoD prison system (White & Young, 2001). As a result of these findings, the Army is currently using AIM for pre-enlistment screening of non-high school graduates and as a training needs diagnostic tool in MOS 95C. Taken collectively, these results indicate that AIM has promise for measuring KSAs important to current and future NCO job performance.

NCO21 Field Test Administration

The AIM was administered to E4 and E5 soldiers in the NCO21 field tests. It was not administered to E6 soldiers because it was intended to serve as a predictor measure and additional data were not necessary to further its development or evaluation.

Field Test Analyses

Table 6.4 shows the internal consistency reliability (Cronbach's alpha) and descriptive statistics for AIM in the NCO21 field test sample. The reliability of the AIM scales was satisfactory and slightly above the reliabilities found in previous research (Young et al., 2000), with one exception. Physical Conditioning had a reliability of .68 in a sample of 21,275 new recruits, as compared with .53 in this NCO sample. One explanation for this difference may be the greater variability of the Physical Conditioning scores of new recruits (Young et al., 2000) as compared with participants in this field test. The AIM also contains a validity index to detect inaccuracies due to socially desirable responding. In the NCO21 field test the level of faking on AIM was low.

Table 6.4. Descriptive Statistics for AIM Scales

AIM Scale	*M*	*SD*	Alpha Reliability
Dependability	1.22	0.24	.72
Adjustment	1.18	0.24	.74
Work Orientation	1.22	0.26	.75
Leadership	1.23	0.29	.76
Agreeableness	1.22	0.28	.73
Physical Conditioning	1.23	0.32	.53

Subgroup analyses of the six AIM scores are shown in Tables 6.5 through 6.10. Previous research has shown small gender differences on the AIM scales (Young et al., 2000). Where differences are found, females tend to score higher than males. In the NCO21 field test, females scored higher than males in Dependability and Agreeableness, a finding consistent with earlier research on new recruits. None of the race effect sizes were statistically significant, a finding also consistent with past AIM research on new recruits.

Mean differences in the AIM scale scores for E4- and E5-level soldiers were examined. Soldiers promoted to E5, a position with greater leadership responsibility than E4, had higher mean scores on Leadership and Dependability.

Mean differences in AIM scale scores as a function of MOS type were also investigated. Most of these comparisons were not statistically significant. Soldiers in combat support and combat service support occupational specialties were higher in Dependability than were soldiers assigned to combat MOS.

Preparation for Validation Research

The field test results indicated that the AIM could be used to reliably measure the intended constructs with some minor modifications. Reliability estimates for five of the six AIM scales ranged from .72 to .76, but were unexpectedly lower for the Physical Conditioning scale. Several items for measuring Physical Conditioning will be added to AIM for the validation effort to improve the internal reliability of this scale.

Table 6.5. Subgroup Differences in AIM Dependability Scores

Group	*n*	*M*	*SD*	Effect Size	*p*
Gender					
Female	61	1.31	0.21	0.40	.004
Male	308	1.21	0.25		
Race					
Black	91	1.26	0.24	0.21	.059
White	222	1.21	0.24		
Pay Grade					
E5	191	1.25	0.22	0.19	.026
E4	187	1.20	0.26		
MOS Type					
Combat Support	95	1.24	0.24	0.30	.034
Combat	85	1.16	0.27		
Combat Service Support	182	1.27	0.22	0.41	<.001
Combat	85	1.16	0.27		
Combat Service Support	182	1.27	0.22	0.13	.316
Combat Support	95	1.24	0.24		

Note. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* referent group. Referent groups (e.g., Whites) are listed second in each pair. Care should be exercised when interpreting the statistics because unequal cell sizes and interaction effects were not taken into account.

Table 6.6. Subgroup Differences in AIM Adjustment Scores

Group	*n*	*M*	*SD*	Effect Size	*p*
Gender					
Female	61	1.13	0.21	-0.24	.088
Male	310	1.19	0.24		
Race					
Black	91	1.19	0.22	0.09	.441
White	224	1.17	0.25		
Pay Grade					
E5	191	1.20	0.24	0.16	.116
E4	189	1.16	0.25		
MOS Type					
Combat Support	95	1.17	0.24	-0.04	.766
Combat	86	1.18	0.25		
Combat Service Support	183	1.19	0.25	0.04	.736
Combat	86	1.18	0.25		
Combat Service Support	183	1.19	0.25	0.09	.482
Combat Support	95	1.17	0.24		

Note. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* referent group. Referent groups (e.g., Whites) are listed second in each pair. Care should be exercised when interpreting the statistics because unequal cell sizes and interaction effects were not taken into account.

Table 6.7. Subgroup Differences in AIM Work Orientation Scores

Group	*n*	*M*	*SD*	Effect Size	*p*
Gender					
Female	61	1.23	0.25	0.08	.710
Male	310	1.21	0.26		
Race					
Black	91	1.21	0.27	-0.04	.768
White	224	1.22	0.26		
Pay Grade					
E5	191	1.25	0.25	0.19	.080
E4	189	1.20	0.27		
MOS Type					
Combat Support	95	1.19	0.25	-0.11	.461
Combat	86	1.22	0.28		
Combat Service Support	183	1.24	0.25	0.07	.548
Combat	86	1.22	0.28		
Combat Service Support	183	1.24	0.25	0.20	.117
Combat Support	95	1.19	0.25		

Note. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* referent group. Referent groups (e.g., Whites) are listed second in each pair. Care should be exercised when interpreting the statistics because unequal cell sizes and interaction effects were not taken into account.

Table 6.8. Subgroup Differences in AIM Leadership Scores

Group	*n*	*M*	*SD*	Effect Size	*p*
Gender					
Female	60	1.20	0.32	-0.14	.333
Male	308	1.24	0.28		
Race					
Black	89	1.20	0.26	-0.17	.138
White	223	1.25	0.30		
Pay Grade					
E5	189	1.26	0.27	0.20	.044
E4	188	1.20	0.30		
MOS Type					
Combat Support	95	1.20	0.30	-0.23	.158
Combat	86	1.26	0.26		
Combat Service Support	180	1.24	0.31	-0.08	.602
Combat	86	1.26	0.26		
Combat Service Support	180	1.24	0.31	0.13	.316
Combat Support	95	1.20	0.30		

Note. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* referent group. Referent groups (e.g., Whites) are listed second in each pair. Care should be exercised when interpreting the statistics because unequal cell sizes and interaction effects were not taken into account.

Table 6.9. Subgroup Differences in AIM Agreeableness Scores

Group	*n*	*M*	*SD*	Effect Size	*p*
Gender					
Female	61	1.30	0.30	0.31	.029
Male	308	1.21	0.28		
Race					
Black	91	1.24	0.29	0.07	.629
White	222	1.22	0.27		
Pay Grade					
E5	191	1.22	0.27	0.00	.993
E4	187	1.22	0.29		
MOS Type					
Combat Support	95	1.23	0.28	0.14	.563
Combat	85	1.19	0.28		
Combat Service Support	182	1.25	0.28	0.21	.098
Combat	85	1.19	0.28		
Combat Service Support	182	1.25	0.28	0.07	.303
Combat Support	95	1.23	0.28		

Note. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* referent group. Referent groups (e.g., Whites) are listed second in each pair. Care should be exercised when interpreting the statistics because unequal cell sizes and interaction effects were not taken into account.

Table 6.10. Subgroup Differences in AIM Physical Conditioning Scores

Group	*n*	*M*	*SD*	Effect Size	*p*
Gender					
Female	59	1.20	0.31	-0.09	.522
Male	307	1.23	0.33		
Race					
Black	87	1.21	0.29	0.04	.766
White	223	1.20	0.32		
Pay Grade					
E5	189	1.25	0.33	0.13	.350
E4	186	1.21	0.31		
MOS Type					
Combat Support	93	1.22	0.32	0.03	.885
Combat	86	1.21	0.35		
Combat Service Support	180	1.25	0.31	0.12	.398
Combat	86	1.21	0.35		
Combat Service Support	180	1.25	0.31	0.09	.475
Combat Support	93	1.22	0.32		

Note. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* referent group. Referent groups (e.g., Whites) are listed second in each pair. Care should be exercised when interpreting the statistics because unequal cell sizes and interaction effects were not taken into account.

Biographical Information Questionnaire (BIQ)

Description and Operational Use

The BIQ measures competencies important to effective NCO performance using rational biodata scales. Self-report biodata scales measure prior behaviors and reactions to specific life events that are indicative of the targeted psychological attributes. Candidate items are reviewed for construct relevance, response variability, readability, non-intrusiveness, and neutrality with respect to social desirability. Response options are scored rationally based on the presumed relationship of the item responses to the underlying psychological construct. The surviving items are pilot tested and revised based on internal consistency reliability and fakability. Previous research has shown that biodata scales can be used to measure personality constructs, have higher criterion-related validity, and are less fakable than traditional self-report personality assessments (Kilcullen, White, Mumford, & Mack, 1995).

The BIQ is actually a compilation of items from existing biodata instruments that have been used by the Army for operational and research purposes. The BIQ taps psychological constructs relevant to leadership, effective performance in uncertain environments, and personal integrity (e.g., Kilcullen, Chen et al., 1999; Kilcullen, Mael, Goodwin, & Zazanis, 1999). The eight constructs measured by the BIQ are also relevant to the NCO21 KSAs described in Chapter 1 (Table 1.3).

Items measuring Hostility to Authority, Manipulativeness, and Social Maturity were drawn from the Army's Assessment of Right Conduct (ARC). These three scales have been related to delinquency criteria and are being used for operational screening and assessment in the Army. In previous research, these attributes have been linked to completion of the Special Forces Assessment and Selection (SFAS) course and a lower incidence of disciplinary infractions among NCO and first term enlisted personnel (e.g., Kilcullen, Mael et al., 1999). Items measuring Tolerance for Ambiguity and Openness were drawn from a biodata instrument that has been used to measure adaptability. In previous research these scales were related to the performance of Special Forces (SF) in Robin Sage, a military exercise consisting of ambiguous and unforeseen dilemmas designed to mimic the SF operational environment (Kilcullen, Chen et al., 1999). In this exercise the team leader's Tolerance for Ambiguity and Openness were primary determinants of the SF team's ability to overcome these challenges and perform successfully. Items for the remaining three biodata scales – Emergent Leadership, Social Perceptiveness, and Interpersonal Skills –were drawn from ARI-sponsored research involving determinants of military and civilian leadership effectiveness. In research with Army civilians these measures, along with individual differences in supervisors' Tolerance for Ambiguity and Openness, were related to effective job performance (Kilcullen, White, Zacarro, & Parker, 2000). Social Perceptiveness and Interpersonal Skills were most important to supervisory performance at lower levels. Tolerance for Ambiguity and Openness were stronger determinants of successful leadership at higher levels of responsibility where the nature of the work is less structured and ill-defined. Definitions for the BIQ scales are shown in Table 6.11.

Table 6.11. BIQ Scale Definitions

Title	Definition
Tolerance for Ambiguity	This scale measures a person's preference for work environments in which the problems (and potential solutions) are unstructured and ill-defined. Those with high tolerance for ambiguity are comfortable working in rapidly changing work environments. Individuals scoring low prefer highly structured and predictable work settings.
Openness	This scale measures the degree to which a person is open to new ideas and experiences. High scorers on this scale are curious, imaginative, have broad interests, and enjoy learning new things. Individuals low in openness dislike extensive thought and contemplation and tend to be set in their ways of doing things.
Hostility to Authority	The degree to which a person respects and is willing to follow legitimate authority figures. High scorers are expressively angered by authority figures and may actively disregard their instructions and policies. Low scorers accept directives from superiors and easily adapt to structured work environments.
Manipulativeness	The degree to which the individual is straightforward and open in his/her interpersonal relationships. Those scoring high in this scale routinely use deception, lies, and short cuts in dealing with others. They are prone to treating others as objects to be used for personal gain and gratification. Low scoring individuals tend to be sincere, aboveboard and straightforward when interacting with others.
Social Maturity	A willingness to follow societal rules and regulations. High scorers tend to be law-abiding and respectful of the rights and property of others. They willingly conform to societal laws, customs, and expectations. Low scorers are highly rebellious and have a history of violating rules and norms.
Social Perceptiveness	This scale measures the degree to which a person can discern and recognize others' emotions and likely behaviors in interpersonal situations. Persons high in social insight are good at understanding others' motives and are less likely to be "caught off guard" by unexpected interpersonal behaviors.
Interpersonal Skill	This scale measures the degree to which a person establishes smooth and effective interpersonal relationships with others. Interpersonally skilled individuals are good listeners, behave diplomatically, and get along well with others. Persons with low scores on this measure have difficulty working with others and may intentionally or unconsciously promote interpersonal conflict and cause hurt feelings.
Emergent Leadership	The scale measures the degree to which a person takes on leadership roles in groups and in his or her interactions with others. High scorers on this scale are looked to for direction and guidance when group decisions are made and readily take on leadership roles.

NCO21 Field Test Administration

The BIQ was administered to E4 and E5 soldiers in the NCO21 field tests. It was not administered to E6 soldiers because it was intended to serve as a predictor measure and additional data was not necessary to further its development or evaluation.

NCO21 Field Test Analyses

The BIQ scales were grouped into three categories reflecting their content and the types of criteria they are used to predict. Table 6.12 presents the internal consistency reliability estimates for the BIQ scales. Six of the eight scales had internal consistency reliability estimates above .70, which is slightly above the estimates found in previous research on biodata scales (Kilcullen et al., 1995; Mumford & Owens, 1987). The reliability estimates for the Tolerance for Ambiguity and Openness scales were lower. The BIQ also contains a Faking Good scale for measuring differences in socially desirable responding. Scores on the Faking Good scale were low, indicating that examinees were generally responding honestly.

Table 6.12. Descriptive Statistics for BIQ Scales

Scale	*M*	*SD*	Alpha Reliability
Information Processing			
Tolerance for Ambiguity	3.23	0.44	0.51
Openness	3.16	0.51	0.62
Personal Integrity			
Hostility to Authority	2.91	0.60	0.75
Manipulativeness	2.37	0.59	0.74
Social Maturity	3.33	0.65	0.73
Interpersonal Characteristics			
Social Perceptiveness	3.58	0.53	0.82
Interpersonal Skill	3.18	0.43	0.71
Emergent Leadership	3.34	0.54	0.81

Note. n = 377-386.

Subgroup analyses for the eight BIQ scores are reported in Tables 6.13 through 6.20. Several statistically significant differences were found, particularly with the Openness and Social Maturity scales. As compared with males, females had significantly higher scores on the Social Maturity scale and lower scores on Hostility to Authority. In contrast, males had a significantly higher mean score on the Openness scale.

Blacks had a significantly higher mean score on Social Maturity and were lower in Hostility to Authority, as compared with Whites. In contrast, as compared with Blacks, Whites had higher mean scores on the Tolerance for Ambiguity and Openness scales. The set of interpersonal scales showed no statistically significant differences due to race.

E5 soldiers were significantly higher than E4s in Tolerance for Ambiguity. E4 soldiers had higher mean scores on the Openness and Social Perceptiveness scales, as compared with E5s. Soldiers in combat support and combat service support occupational specialties had higher mean scores on the Social Maturity and Interpersonal Skill scales, as compared with soldiers in combat MOS. In addition, soldiers assigned to combat service support MOS were significantly lower in Hostility to Authority than were soldiers in combat or combat support occupations.

Table 6.13. Subgroup Differences in BIQ Tolerance for Ambiguity Scores

Group	*n*	*M*	*SD*	Effect Size	*p*
Gender					
Female	61	3.19	0.46	-0.12	.394
Male	304	3.24	0.44		
Race					
Black	86	3.17	0.43	-0.28	.042
White	222	3.29	0.43		
Pay Grade					
E5	193	3.28	0.44	0.21	.034
E4	179	3.19	0.43		
MOS Type					
Combat Support	92	3.20	0.37	-0.04	.699
Combat	88	3.22	0.46		
Combat Service Support	178	3.26	0.47	0.09	.508
Combat	88	3.22	0.46		
Combat Service Support	178	3.26	0.47	0.16	.255
Combat Support	92	3.20	0.37		

Note. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* referent group. Referent groups (e.g., Whites) are listed second in each pair. Care should be exercised when interpreting the statistics because unequal cell sizes and interaction effects were not taken into account.

Table 6.14. Subgroup Differences in BIQ Openness Scores

Group	*n*	*M*	*SD*	Effect Size	*p*
Gender					
Female	61	3.01	0.50	-0.33	.018
Male	308	3.18	0.51		
Race					
Black	86	3.06	0.54	-0.27	.036
White	226	3.20	0.51		
Pay Grade					
E5	193	3.11	0.51	-0.20	.048
E4	183	3.21	0.51		
MOS Type					
Combat Support	92	3.21	0.52	-0.02	.891
Combat	88	3.22	0.47		
Combat Service Support	182	3.11	0.52	-0.23	.096
Combat	88	3.22	0.47		
Combat Service Support	182	3.11	0.52	-0.19	.138
Combat Support	92	3.21	0.52		

Note. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* referent group. Referent groups (e.g., Whites) are listed second in each pair. Care should be exercised when interpreting the statistics because unequal cell sizes and interaction effects were not taken into account.

Table 6.15. Subgroup Differences in BIQ Hostility to Authority Scores

Group	*n*	*M*	*SD*	Effect Size	*p*
Gender					
Female	61	2.69	0.56	-0.47	<.001
Male	303	2.96	0.59		
Race					
Black	86	2.80	0.56	-0.27	.045
White	222	2.96	0.60		
Pay Grade					
E5	192	2.87	0.60	-0.17	.098
E4	179	2.97	0.58		
MOS Type					
Combat Support	92	2.97	0.62	-0.09	.469
Combat	88	3.03	0.64		
Combat Service Support	177	2.80	0.54	-0.36	.002
Combat	88	3.03	0.64		
Combat Service Support	177	2.80	0.54	-0.27	.028
Combat Support	92	2.97	0.62		

Note. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* referent group. Referent groups (e.g., Whites) are listed second in each pair. Care should be exercised when interpreting the statistics because unequal cell sizes and interaction effects were not taken into account.

Table 6.16. Subgroup Differences in BIQ Manipulativeness Scores

Group	*n*	*M*	*SD*	Effect Size	*p*
Gender					
Female	61	2.26	0.57	-0.24	.088
Male	302	2.40	0.59		
Race					
Black	86	2.37	0.60	0.03	.809
White	221	2.35	0.56		
Pay Grade					
E5	192	2.31	0.56	-0.18	.060
E4	178	2.42	0.61		
MOS Type					
Combat Support	91	2.38	0.60	-0.15	.276
Combat	88	2.48	0.65		
Combat Service Support	177	2.29	0.56	-0.29	.015
Combat	88	2.48	0.65		
Combat Service Support	177	2.29	0.56	-0.15	.240
Combat Support	91	2.38	0.60		

Note. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* referent group. Referent groups (e.g., Whites) are listed second in each pair. Care should be exercised when interpreting the statistics because unequal cell sizes and interaction effects were not taken into account.

Table 6.17. Subgroup Differences in BIQ Social Maturity Scores

Group	*n*	*M*	*SD*	Effect Size	*p*
Gender					
Female	61	3.73	0.53	0.74	<.001
Male	304	3.25	0.64		
Race					
Black	86	3.55	0.61	0.47	<.001
White	222	3.25	0.64		
Pay Grade					
E5	193	3.39	0.59	0.17	.071
E4	179	3.27	0.69		
MOS Type					
Combat Support	92	3.32	0.68	0.33	.022
Combat	88	3.08	0.72		
Combat Service Support	178	3.46	0.57	0.53	<.001
Combat	88	3.08	0.72		
Combat Service Support	178	3.46	0.57	0.21	.071
Combat Support	92	3.32	0.68		

Note. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* referent group. Referent groups (e.g., Whites) are listed second in each pair. Care should be exercised when interpreting the statistics because unequal cell sizes and interaction effects were not taken into account.

Table 6.18. Subgroup Differences in BIQ Social Perceptiveness Scores

Group	*n*	*M*	*SD*	Effect Size	*p*
Gender					
Female	61	3.48	0.47	-0.20	.115
Male	300	3.59	0.55		
Race					
Black	86	3.58	0.51	-0.04	.733
White	219	3.60	0.53		
Pay Grade					
E5	191	3.52	0.53	-0.25	.029
E4	177	3.65	0.53		
MOS Type					
Combat Support	89	3.58	0.52	0.05	.728
Combat	88	3.55	0.58		
Combat Service Support	177	3.60	0.53	0.09	.534
Combat	88	3.55	0.58		
Combat Service Support	177	3.60	0.53	0.04	.821
Combat Support	89	3.58	0.52		

Note. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* referent group. Referent groups (e.g., Whites) are listed second in each pair. Care should be exercised when interpreting the statistics because unequal cell sizes and interaction effects were not taken into account.

Table 6.19. Subgroup Differences in BIQ Interpersonal Skill Scores

Group	*n*	*M*	*SD*	Effect Size	*p*
Gender					
Female	61	3.21	0.40	0.09	.526
Male	304	3.17	0.43		
Race					
Black	86	3.20	0.43	0.08	.534
White	220	3.17	0.42		
Pay Grade					
E5	192	3.20	0.42	0.07	.525
E4	177	3.17	0.43		
MOS Type					
Combat Support	90	3.21	0.43	0.33	.027
Combat	88	3.07	0.43		
Combat Service Support	177	3.23	0.41	0.37	.004
Combat	88	3.07	0.43		
Combat Service Support	177	3.23	0.41	0.05	.770
Combat Support	90	3.21	0.43		

Note. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* referent group. Referent groups (e.g., Whites) are listed second in each pair. Care should be exercised when interpreting the statistics because unequal cell sizes and interaction effects were not taken into account.

Table 6.20. Subgroup Differences in BIQ Emergent Leadership Scores

Group	*n*	*M*	*SD*	Effect Size	*p*
Gender					
Female	61	3.22	0.51	-0.27	.053
Male	309	3.37	0.55		
Race					
Black	86	3.32	0.56	-0.09	.459
White	226	3.37	0.54		
Pay Grade					
E5	193	3.38	0.51	0.12	.230
E4	184	3.31	0.58		
MOS Type					
Combat Support	92	3.30	0.53	-0.04	.741
Combat	88	3.32	0.51		
Combat Service Support	183	3.38	0.56	0.12	.439
Combat	88	3.32	0.51		
Combat Service Support	183	3.38	0.56	0.15	.253
Combat Support	92	3.30	0.53		

Note. Effect sizes calculated as (mean of non-referent group – mean of referent group)/*SD* referent group. Referent groups (e.g., Whites) are listed second in each pair. Care should be exercised when interpreting the statistics because unequal cell sizes and interaction effects were not taken into account.

Tables 6.21 and 6.22 show intercorrelations among the operational predictor scores described in this chapter for E4 and E5 soldiers, respectively. The correlations show a sensible pattern. For example, the temperament-related scores from the AIM and BIQ are generally uncorrelated or minimally correlated with the cognitive scores from the ASVAB, particularly at the E4 level. In contrast, they show consistent moderate correlations with each other. The pattern of correlations is very similar across the two grades, with the notable exception of the AIM Physical Conditioning scale.

Preparation for Validation Research

The field test results indicated that the BIQ could be used to provide reliable measures of the intended constructs with only a few minor changes. Several items for measuring Openness and Tolerance for Ambiguity will be added for the validation to improve the internal reliability of these scales.

Table 6.21. Correlations Among Non-Experimental Measures for E4 Soldiers

	AFQT	GT	AIM Dep	AIM Adj	AIM Work	AIM Lead	AIM Agree	AIM Phys	BIQ Host	BIQ Man	BIQ Perc	BIQ Mat	BIQ Tol	BIQ Open	BIQ Lead
AFQT															
GT (self-report)	.39*														
AIM															
Dependability	.14	.02													
Adjustment	.04	-.06	.44*												
Work Orientation	.06	.10	.49*	.41*											
Leadership	.05	.12	.27*	.43*	.63*										
Agreeableness	.08	-.01	.65*	.50*	.40*	.10									
Physical Conditioning	.07	.04	.32*	.28*	.46*	.13	.37*								
BIQ															
Hostility to Authority	-.13	-.04	-.64*	-.34*	-.32*	-.12*	-.56*	-.30*							
Manipulativeness	-.10	-.03	-.63*	-.32*	-.43*	-.27*	-.50*	-.33*	.58*						
Social Perceptive	.14	.01	-.00	.19*	.20*	.30*	-.06	-.02	.10	-.14					
Social Maturity	.12	-.03	.68*	.20*	.21*	-.00	.60*	.23*	-.69*	-.60*	-.14				
Toler. for Ambig.	.16*	.07	.28*	.31*	.34*	.33*	.32*	.15*	-.24*	-.41*	.10	.21*			
Openness	.16*	.09	-.09	-.07	.13	.31*	-.19*	.00	.15*	.04	.36*	-.15	.04		
Emergent Leadership	.11	.11	.12	.28*	.42*	.61*	-.05	.03	-.02	-.10	.53*	-.13	.14	.29*	
Interpersonal Skills	.26*	.10	.57*	.39*	.43*	.30*	.54*	.27*	-.65*	-.61*	.21*	.53*	.37*	.07	.28*

Note. n = 169-187.
*$p < .05$.

Table 6.22. Correlations Among Non-Experimental Measures for E5 Soldiers

	AFQT	GT	AIM Dep	AIM Adj	AIM Work	AIM Lead	AIM Agree	AIM Phys	BIQ Host	BIQ Man	BIQ Perc	BIQ Mat	BIQ Tol	BIQ Open	BIQ Lead
AFQT															
GT (self-report)	.56*														
AIM															
Dependability	.12	.03													
Adjustment	.04	.06	.45*												
Work Orientation	.03	.00	.46*	.42*											
Leadership	.12	.12	.39*	.43*	.71*										
Agreeableness	.00	-.06	.57*	.52*	.38*	.18*									
Physical Conditioning	-.15*	-.06	.15*	.18*	.28*	.01	.19*								
BIQ															
Hostility to Authority	-.04	-.05	-.54*	-.40*	-.30*	-.28*	-.50*	-.09							
Manipulativeness	-.16*	-.18*	-.48*	-.40*	-.41*	-.35*	-.41*	-.08	.55*						
Social Perceptive	.15*	.11	.09	.22*	.23*	.31*	.01	-.07	.07	-.14					
Social Maturity	.08	.09	.53*	.31*	.22*	.12	.47*	.15*	-.61*	-.49*	-.09				
Toler. for Ambig.	.13	.05	.20*	.33*	.40*	.42*	.18*	-.00	-.28*	-.47*	.19*	.18*			
Openness	.37*	.15*	.02	-.02	-.06	-.03	-.02	-.16*	.14	.09	.32*	-.09	-.06		
Emergent Leadership	.15*	.16*	.17*	.24*	.40*	.55*	.03	-.07	.04	-.14	.65*	-.13	.17*	.24*	
Interpersonal Skills	.15*	.16*	.53*	.51*	.32*	.29*	.55*	-.02	-.64*	-.53*	.16*	.50*	.36*	.12	.22*

Note. n = 181-191.
*$p < .05$.

CHAPTER 7: CROSS-INSTRUMENT ANALYSES

Overview

This chapter provides a preliminary assessment of the interrelationships among, and criterion-related validity of, the NCO21 predictor scores. The analyses of interrelationships included simple, uncorrected bivariate correlations among the predictors. Ideally, this type of analysis would yield evidence to support the construct validity of these measures. Given the small sample sizes and heterogeneity of the instruments, however, it would be difficult to offer conclusive evidence in this regard. Instead, this chapter will present preliminary results related to the establishment of construct validity – an exploratory, descriptive assessment of the interrelationships between the predictors, highlighting the results of greatest interest. More complete evaluations of construct validity will be conducted with the criterion-related validation data.

Similarly, the data are not substantial enough to support strong assertions about criterion-related validity. Thus, this chapter presents preliminary evidence of the criterion-related validity of the predictors from zero-order correlations between the predictors and the performance criteria. Correlations were corrected for unreliability in the criterion measures. Significance tests were performed; however, the sample sizes on which they are based are relatively small.

All analyses were conducted by grade (i.e., E4, E5, and E6), as the measures were differentially administered by grade. Table 7.1 presents the predictor and criterion instruments administered to soldiers of each grade. Although data will be presented for both the experimental (i.e., SJT, ExAct, PFF21, and interview) and non-experimental (i.e., ASVAB, AIM, BIQ) measures, this chapter will primarily focus on the preliminary correlations among, and criterion-related validity estimates for, the experimental measures.

Table 7.1. Predictor and Criterion Measures Administered to Soldiers by Grade

Instrument	Grade		
	E4	E5	E6
Situational Judgment Test (SJT)	✓	✓	✓
Experience and Activities Record (ExAct)	✓	✓	✓
Personnel File Form-21 (PFF21)	✓	✓	✓
Assessment of Individual Motivation (AIM)	✓	✓	
Biographical Information Questionnaire (BIQ)	✓	✓	
Semi-Structured Interview	✓	✓	
ASVAB AFQT Composite [from EMF]	✓	✓	✓
ASVAB GT Composite [self-report]	✓	✓	✓
Observed Performance Rating Scales		✓	✓
Expected Future Performance Rating Scales		✓	✓

Covariance of Predictor Scores

Preliminary evidence of construct validity was examined through correlations among predictor scores within each grade. To maximize the sample size underlying each correlation, pairwise deletion of missing scores was used. In this section, correlations among the

experimental predictor scores are presented, followed by correlations between the scores on the experimental and non-experimental measures.

Correlations Among Experimental Predictors

Table 7.2 shows the bivariate correlations among the scores on the four experimental predictor measures. A number of significant relationships were found. The most salient results are described below.

Situational Judgment Test

The field test version of the SJT had two forms that targeted different KSAs with some overlap. Both forms covered Directing, Monitoring, and Supervising Individual Subordinates and Training Others. In addition, Form A measured (a) Relating to and Supporting Peers; (b) Cultural Tolerance; and (c) Motivating, Leading, and Supporting Individual Subordinates. Form B also assessed (a) Concern for Soldiers' Quality of Life, (b) Problem-Solving/Decision Making Skill, and (c) Team Leadership. This difference in content focus at least partly accounts for the different patterns of correlations that were found between the two SJT forms.

At the E4 level, SJT (Form A) correlated significantly with the ExAct Computer score, PFF Awards, and PFF Civilian Education. It correlated significantly with PFF Awards and PFF Military Education at the E6 level and did not correlate with any scores at the E5 level. For E4 soldiers, SJT (Form B) correlated significantly with the ExAct General score and PFF Memoranda/Letters. For E5 soldiers, it correlated significantly with PFF PPW Achievement, and for E6 soldiers the SJT (Form B) was significantly related to PFF Certificates. Although the SJT was not significantly correlated with the interview, the correlations were positive and of a similar magnitude as the uncorrected correlations found in the ECQUIP project between a situational judgment test (i.e., Army Leadership Questionnaire) and an interview (Peterson et al., 1997).

Semi-Structured Interview

In general, the interview correlations with other predictors were stronger at the E4 level than at the E5 level, but the interview score was correlated with different variables for each grade. This might suggest that either questions associated with different KSAs were easier to answer for soldiers of different grades or some areas in the interview differentiated among soldiers more than others at the two grades. In particular, the interview composite was significantly correlated with the ExAct scales (both General and Computer) and PFF Awards (unweighted and weighted) for E4 soldiers. For E5 soldiers, there was a significant positive relationship between the interview composite and Army Physical Fitness Test (APFT; weighted score) and the interview score was negatively correlated with Disciplinary Actions. This may suggest that E5 interviews were more heavily influenced than the E4 interviews by the Integrity and Discipline or Military Presence KSAs. The different patterns of correlations across grades may also be due, in part, to differential MOS sampling across grades. Although both grades were primarily represented by combat service support MOS, nearly one-third (31%) of E5 soldiers interviewed were in combat support MOS (as opposed to 16% for E4s). In contrast, 28% of E4 interviewees were in combat MOS compared to 15% of E5s.

Table 7.2. Correlations Among Experimental Predictor Measures

	SJT-A	SJT-B	Interview	ExAct-CPT	ExAct-GEN
E4 Soldiers					
Interview Composite[a]	.17	.14			
ExAct Computer	.23*	.05	.37*		
ExAct General	.13	.23*	.21*		
PFF Awards	.23*	-.01	.24*	.03	.39*
PFF Awards (wt)	.18	-.05	.24*	.03	.36*
PFF PPW Achievement	.05	.15	.05	.14	.28*
PFF Memoranda/Letters	.14	.27*	.03	-.00	.36*
PFF Achievement Certificates	-.04	.15	-.11	.06	.24*
PFF PPW Military Education	-.18	.03	.21	.11	.03
PFF PPW Civilian Education	.29*	-.10	.19	.13	-.14
PFF Disciplinary Actions	.02	-.12	-.10	-.04	.08
PFF APFT score	-.07	.09	.00	-.08	.06
PFF APFT score (wt)	-.07	.05	.07	-.05	.15*
PFF Weapons Qualification	.07	.04	.20	.03	.32*
PFF Military Training	-.01	.02	.10	-.05	.29*
E5 Soldiers					
Interview Composite[a]	.12	.07			
ExAct Computer	.06	.09	.08		
ExAct General	.04	-.03	.19		
PFF Awards	.02	-.08	.06	.09	.40*
PFF Awards (wt)	.01	-.15	.05	.07	.34*
PFF PPW Achievement	-.13	.20*	.05	-.05	.19*
PFF Memoranda/Letters	-.02	-.15	-.09	-.02	.31*
PFF Achievement Certificates	-.16	-.04	.02	-.05	.35*
PFF PPW Military Education	-.07	.11	.04	.10	.18*
PFF PPW Civilian Education	.10	.18	.05	.02	-.14
PFF Disciplinary Actions	-.07	.06	-.31*	-.02	.07
PFF APFT score	-.13	-.05	.15	.09	.13
PFF APFT score (wt)	-.13	-.14	.23*	.06	.10
PFF Weapons Qualification	.00	.09	-.02	-.00	.36*
PFF Military Training	-.09	-.03	.19	.02	.30*
E6 Soldiers					
ExAct Computer	.01	-.03	--		
ExAct General	.13	.20	--		
PFF Awards	.31*	.25	--	.04	.35*
PFF Awards (wt)	.27	.26	--	.09	.27*
PFF PPW Achievement	.18	.24	--	.02	.19
PFF Memoranda/Letters	-.09	.24	--	-.11	.14
PFF Achievement Certificates	.07	.28*	--	.03	.25*
PFF PPW Military Education	.37*	.18	--	-.08	.18
PFF PPW Civilian Education	.21	.23	--	.20*	.12
PFF Disciplinary Actions	-.10	.03	--	-.14	-.05
PFF APFT score	-.12	-.07	--	-.03	.11
PFF APFT score (wt)	-.16	.02	--	-.06	.12
PFF Weapons Qualification	.19	.10	--	-.10	.13
PFF Military Training	.04	.09	--	-.13	.17

Note. n_{E4} = 84-204, n_{E5} = 95-210, n_{E6} = 46-97. Correlations between scores from the same instrument are presented in previous chapters. wt = weighted.

[a]Interview composite score does not include the MOS/Occupation-Specific Knowledge and Skill score to maximize sample size. n = 46-105.

*$p < .05$.

Experience and Activities Record and Personnel File Form 21

One or more PFF21 scores were significantly correlated with at least one score from each of the other experimental measures across grades, with the relationships stronger at the E4 and E5 grades compared to E6. Because the ExAct and PFF21 are both self-report instruments measuring biodata or archival information, one would expect to find some relationships between the scores on these measures. As expected, ExAct General scores were correlated with a number of PFF21 scales across grades. As mentioned, ExAct General scores were also correlated with interview scores for E4 soldiers. It is unclear why a similar relationship was not found with E5 soldiers, as the level of variance in the scores on the ExAct and interview did not differ appreciably across grades. In addition, ExAct Computer scores correlated significantly with PFF Civilian Education for E6 soldiers; however, the Computer score was not correlated significantly with any of the other experimental predictors for the various grades.

Summary

In general, the overlap in the assessment of several KSAs across multiple predictor measures would suggest some of the NCO21 predictor scores should show low to moderate relationships with each other. Thus, the presence of significant correlations between the experimental predictor measures is consistent with expectations. Further evaluations of covariation in the predictor scores are presented in the correlations between the experimental and non-experimental predictors.

Correlations Between Experimental and Nonexperimental Predictors

Tables 7.3-7.5 present the uncorrected bivariate correlations between the experimental and non-experimental predictors for E4 - E6 soldiers, respectively. The tables contain a number of significant relationships. In particular, scores on the AIM and BIQ were correlated with all four experimental predictor measures for E4 and E5 soldiers.

Situational Judgment Test

SJT (Form A) correlated significantly with AFQT and GT for E4 soldiers, and Form B was correlated with AFQT and GT for E5 soldiers. No significant relationships were found between either SJT form and the two ASVAB composite scores for E6 soldiers, but this is likely due to low sample size. The significant correlations with AFQT and GT support previous research with SJTs. For example, the McDaniel et al. (2001) meta-analysis found that, when corrected for range restriction, SJTs correlated .42 with general intelligence.

At the E4 level, SJT (Form A) correlated significantly with AIM Dependability and Agreeableness scores. It did not correlate significantly with any AIM scales at the E5 level. SJT (Form B), however, was correlated significantly with AIM Physical Conditioning at the E4 level and AIM Dependability at the E5 level. With regard to the BIQ, SJT (Form A) was correlated significantly (in the expected directions) with Hostility to Authority, Manipulativeness, Social Maturity, and Interpersonal Skills at the E4 level. SJT (Form A) did not correlate with any BIQ scales at the E5 level. For E4 soldiers, SJT (Form B) correlated significantly with BIQ Openness and Interpersonal Skills. At the E5 level, it correlated significantly with Social Maturity and Interpersonal Skills.

Table 7.3. Correlations Between Experimental and Non-Experimental Measures for E4 Soldiers

Non-Experimental Measure	SJT[a] A	SJT[a] B	INT[a]	ExAct CPT	ExAct GEN	PFF Award	PFF Award (wt)	PFF PPW Ach	PFF Memo	PFF Cert	PFF[a] Mil Ed	PFF Civ Ed	PFF Disc Actn	PFF AP FT	PFF APFT (wt)	PFF Weapon Qualif	PFF Mil Train
AFQT	.35*	.19	.16	.24*	-.02	-.03	-.07	.07	.05	.14	-.00	.35*	-.08	-.12	-.11	.03	-.03
GT (self-report)	.26*	.14	.10	.22*	.14	-.04	-.09	.17*	.09	13	-.10	.11	.03	.04	.06	.09	10
AIM																	
Dependability	.26*	.20	.33*	.13	-.01	-.02	.00	.01	-.02	-.06	.06	.23*	-.18*	.02	-.04	-.10	-.07
Adjustment	.11	.09	.22*	.10	.16*	.04	.05	.07	.13	-.03	-.05	.11	-.07	-.03	-.01	-.02	-.01
Work Orientation	.14	.13	.41*	.19*	.31*	.01	.02	.12	.13	-.01	-.04	.08	-.10	.01	.03	.02	.02
Leadership	.13	.08	.35*	.23*	.50*	.13	.12	.23*	.20*	.10	-.17	-.02	.08	.00	.06	.11	.10
Agreeableness	.33*	.16	.22*	.03	-.17*	.05	.07	-.08	-.08	-.17*	.12	.26*	-.10	.10	.01	-.13	-.08
Physical Cond.	.06	.23*	.23*	-.07	-.06	.02	.01	-.02	.02	-.08	.13	.03	-.19*	.13	.19*	-.03	.04
BIQ																	
Hostility to Auth.	-.33*	-.18	-.14	-.05	.21*	.08	.09	-.01	.11	.10	-.09	-.17	.12	.02	.08	.16*	.18*
Manipulativeness	-.23*	-.15	-.23	-.07	.04	-.02	-.01	-.04	-.05	.02	-.02	-.12	.13	.04	.07	.10	.13
Social Perceptive	.05	.13	.15	.24*	.32*	.10	.10	.00	.02	.02	-.08	-.10	-.21	.03	.05	.12	.05
Social Maturity	.26*	.15	.21	.13	-.24*	-.09	-.09	-.08	-.09	-.12	.14	.25*	-.09	-.01	-.10	-.21*	-.21
Toler. for Ambig.	.14	.19	.10	.05	-.00	.03	.05	.07	-.04	.05	.01	.03	.06	-.01	-.00	-.02	-.01
Openness	.09	.33*	.03	.20*	.29*	.08	.03	-.00	.17	.06	-.10	-.13	-.03	.03	.10	.21*	.16*
Emergent Leadership	.13	.04	.19	.29*	.57*	.18*	.16*	.17*	.16*	.10	-.06	-.05	.09	-.04	.07	.09	.05
Interpersonal Skills	.41*	.31*	.26*	.21*	.09	-.01	-.06	.02	-.04	-.03	.11	.16*	-.06	.06	.00	-.12	-.10

Note. n = 153-195. [a]n = 81-106. wt = weighted.
*$p < .05$.

Table 7.4. Correlations Between Experimental and Non-Experimental Measures for E5 Soldiers

Non-Experimental Measure	SJT A	SJT B	INT	ExAct CPT	ExAct GEN	PFF Award	PFF Award (wt)	PFF PPW Ach	PFF Memo	PFF Cert	PFF Mil Ed	PFF Civ Ed	PFF Disc Actn	PFF AP FT	PFF APFT (wt)	PFF Weapon Qualif	PFF Mil Train
AFQT	.09	.31*	.02	.16	-.03	-.16*	-.19*	-.09	-.14*	-.12	-.08	.19	.11	-.14	-.18*	-.07	-.18*
GT (self-report)	.09	.27*	-.08	.09	.11	-.04	-.09	.01	-.00	.02	.03	.29*	-.01	-.07	-.07	.04	-.03
AIM																	
Dependability	.19	.36*	.07	.24*	.01	-.10	-.09	.05	.03	-.04	.01	.13	-.05	-.11	-.07	-.11	-.14
Adjustment	-.01	.12	.07	.17*	.09	.04	.01	.12	.16*	.08	.10	-.02	-.00	-.02	.09	.04	.10
Work Orientation	.12	.18	.30*	.21*	.26*	.13	.07	.17*	.00	.08	.07	.04	-.09	.18*	.23*	.05	.20*
Leadership	.07	.05	.25*	.16*	.30*	.03	-.01	.14	-.01	.08	-.01	-.08	-.10	.10	.15*	.15*	.21*
Agreeableness	.04	.20	.15	.12	-.10	-.06	-.04	.11	.09	.04	.08	.11	-.07	.04	.08	-.11	.00
Physical Cond.	.06	.09	.23*	.10	-.07	.01	.03	-.02	.00	-.06	.08	.08	-.11	.19*	.39*	-.12	.17
BIQ																	
Hostility to Auth.	-.04	-.18	-.21*	-.06	.12	.08	.09	-.13	-.04	-.07	-.07	-.14	-.03	.05	-.00	.11	.03
Manipulativeness	-.15	-.20	-.26*	-.02	-.02	.05	.08	-.15*	-.04	-.10	-.07	-.07	.01	.13	.08	.00	.03
Social Perceptive	.13	.02	.05	.18*	.28*	.09	.06	-.10	.09	.16*	-.08	-.13	.07	.06	.03	.15*	.12
Social Maturity	.17	.27*	.18	.13	-.12	-.01	-.02	.10	.06	-.06	.10	.18*	-.05	-.14	-.05	-.15*	-.10
Toler. for Ambig.	.04	.16	.34*	.08	.11	.06	-.00	.15*	.04	.14	.06	-.01	-.00	.02	-.00	.08	.07
Openness	.04	.05	-.10	.19	.09	-.07	-.11	-.20*	-.16*	-.12	.06	.06	.10	-.02	-.04	.03	-.02
Emergent Leadership	.20	.09	.16	.23*	.40*	.13	.05	.02	.03	.13	.00	-.04	.01	.13	.13	.10	.12
Interpersonal Skills	.07	.37*	.21*	.13	-.05	.01	-.06	.09	.01	.06	.08	.03	.04	.00	.02	-.02	.04

Note. $n = 179 - 200$. [a]$n = 90\text{-}102$. wt = weighted.
*$p < .05$.

Table 7.5. Correlations Between Scores on Experimental Predictor Measures and ASVAB Composites for E6 Soldiers

Experimental Measures	AFQT	GT
SJT Form A	.11	.06
SJT Form B	.06	.14
ExAct Computer	.03	.04
ExAct General	-.04	.17
PFF Awards	.17	.18
PFF Awards (wt)	.05	.13
PFF PPW Achievement	.06	.10
PFF Memoranda/Letters	-.06	.07
PFF Achievement Certificates	-.17	-.06
PFF PPW Military Education	.22*	.18
PFF PPW Civilian Education	.16	.19
PFF Disciplinary Actions	-.10	-.02
PFF APFT score	.02	.09
PFF APFT score (wt)	.00	.11
PFF Weapons Qualification	.00	.05
PFF Military Training	.06	.15

Note. n = 44-95. wt = weighted.
*$p < .05$.

These significant correlations align with results found in the ECQUIP project. In that research, the Army Leadership Questionnaire (ALQ) SJT was correlated .08 to .15 with the dimensions of the ABLE (Assessment of Background and Life Experiences; the predecessor to the AIM) similar to those of the AIM. The correlations in the current effort appear to be either similar or slightly higher than those found in ECQUIP, particularly for E4 soldiers.

Semi-Structured Interview

Considering that the interview assesses two types of communications/interpersonal relations KSAs (Oral Communication Skill; Relating to and Supporting Peers), one would expect a positive relationship between the interview and the BIQ measure of Interpersonal Relations. As expected, the interview composite score correlated significantly with the BIQ Interpersonal Skills scale score for both E4 and E5 soldiers. Similarly, the results showed significant relationships between the interview composite and all of the AIM scores for E4 soldiers. At the E5 level, only correlations with AIM Work Orientation, Leadership, and Physical Conditioning were significant. This makes sense because the interview measures constructs similar to these AIM scales (i.e., Leadership Skill/Potential, Level of Effort and Initiative on the Job). For E5 soldiers, BIQ Hostility to Authority and BIQ Manipulativeness scores were negatively associated with the interview composite score and the interview was positively associated with Tolerance for Ambiguity.

The interview was not significantly correlated with AFQT or GT at the E4 or E5 levels. The .16 correlation between the interview and AFQT for E4 soldiers is similar to that found in the ECQUIP project (uncorrected r = .12). One possible reason for this lack of significant

relationship is that AFQT and GT scores measure cognitive ability, or "can-do" aspects of performance, and the interview aims to also assess "will-do" or "have-done" types of performance, which can be influenced by other factors such as motivation.

Experience and Activities Record

The ExAct scale scores correlated significantly with a number of AIM and BIQ scores. Specifically, both ExAct scores correlated with AIM Work Orientation and Leadership for E4 and E5 soldiers. In addition, among E4 soldiers, the ExAct General score correlated with AIM Adjustment and Agreeableness. The ExAct Computer score correlated significantly with AIM Dependability and Adjustment for E5 soldiers. Further, ExAct General scores correlated with all of the BIQ scales for E4 soldiers with the exception of three (Manipulativeness, Tolerance for Ambiguity, and Interpersonal Skills). Both ExAct scale scores correlated with BIQ Social Perceptiveness and BIQ Emergent Leadership for E5 soldiers. Further, ExAct Computer scores correlated significantly with AFQT and GT for E4 soldiers. No significant relationships were found between ExAct and AFQT or GT for E5 or E6 soldiers.

Personnel File Form 21

Given the similarity in constructs, a significant positive relationship was expected between PFF APFT scores and AIM Physical Conditioning. Consistent with prediction, there was a significant positive correlation between the weighted PFF APFT score and AIM Physical Conditioning for both E4 and E5 soldiers.

In addition, the PFF21 scores correlated significantly with several other AIM scores for E4 soldiers, including Dependability, Leadership, and Agreeableness. E5 soldier correlations revealed significant relationships between PFF21 scales and AIM Work Orientation, Adjustment, and Leadership. In relation to the BIQ, some of the PFF21 scores correlated significantly with Emergent Leadership, Social Maturity, Interpersonal Skills, Openness, and Hostility to Authority for E4 soldiers. At the E5 level, various PFF scores correlated significantly with BIQ Manipulativeness, Tolerance for Ambiguity, Openness, Social Perceptiveness, and Social Maturity. For the most part, correlations between PFF21 scores and BIQ scores were in the expected direction (e.g., negative correlations with unfavorable constructs such as Manipulativeness and positive correlations with positive-oriented BIQ scales). However, high scores on Weapons Qualifications and on Military Training showed significant positive relationships with Hostility to Authority for E4 soldiers. In addition to the presence of significant relationships with the AIM and BIQ scores, at least one PFF21 score correlated significantly with AFQT for each rank. GT correlated with different PFF21 scores for E4 and E5 soldiers but not with any of the PFF21 scores for E6 soldiers.

Summary

The intercorrelations between the new and existing predictor measures showed no unexpected patterns. In particular, there are no high correlations that suggest unnecessary redundancy across the instruments.

Criterion-Related Validity of Predictor Scores

Preliminary evidence of criterion-related validity was examined by computing zero-order correlations between predictors and the two composite criterion scores for each applicable pay grade (i.e., E5 and E6). To maximize the sample size underlying each correlation, pairwise deletion of missing scores was used to calculate the values in each correlation matrix. As indicated in Chapter 2, both the Observed Performance Rating Scale and the Future Performance Rating Scale composite scores demonstrated low interrater reliability. Thus, the correlations between the predictor and composite criterion scores have been corrected for unreliability in the criterion using the following formula (Crocker & Algina, 1986, p. 237):

$$r_{y_T x} = \frac{r_{yx}}{\sqrt{r_{yy}}}$$

where r_{yx} is the observed correlation coefficient between the predictor and criterion composite, r_{yy} is the weighted interrater reliability estimate, and $r_{y_T x}$ is the observed correlation corrected for unreliability in the criterion. The single rater reliability and the reliability of the mean of two raters (via the Spearman-Brown formula) were estimated using subsamples (Observed Performance, $n = 72$; Expected Future Performance, $n = 71$) of soldiers who were each rated by two supervisors.[15] Obviously, a reliability coefficient cannot be estimated on subsamples (Observed and Expected Future Performance, $n = 139$) of soldiers who had only one rater. The interrater reliability estimate was weighted to take into account the different sample sizes of soldiers for whom the criterion composite scores were based on either (a) the mean of a single supervisor's ratings (Observed and Expected Future Performance, $n = 139$) or (b) the mean of two supervisors' ratings (Observed Performance, $n = 72$; Expected Future Performance, n = 71) averaged over the scales in each composite. The formula used to calculate the weighted interrater reliability estimate was

$$r_{yy}^{w} = \frac{z_{IRR1} n_1 + z_{IRR2} n_2}{n_{total}}$$

Specifically, this weighted estimate was calculated by (a) taking the r to z transformation of the inter-rater reliability coefficient for a single rater (z_{IRR1}) and for two raters (z_{IRR2}, estimated using the Spearman Brown formula), (b) multiplying these transformations by their respective sample sizes, (c) summing the two products, (d) dividing the sum by total sample size, and (e) retransforming the weighted z to r. Table 7.6 presents the corrected correlations between the predictor and criterion scores.

[15] The single rater reliability estimates used here are for soldiers who were each rated by at least two supervisors; $r_{IRR1} = .34$ for the Observed Performance Rating Scales composite and $r_{IRR1} = .16$ for the Expected Future Performance Rating Scales composite (see Tables 2.5 and 2.11, respectively).

Table 7.6. Uncorrected and Corrected Correlations Between Predictors and Criteria for E5 and E6 Soldiers

Predictor	Observed Performance Composite				Expected Future Performance Composite			
	Uncorrected		Corrected[b]		Uncorrected		Corrected	
	E5	E6	E5	E6	E5	E6	E5	E6
SJT Form A	-.04	.24	-.07	.35	.00	.11	.00	.24
SJT Form B	-.02	-.03	-.02	-.05	.13	-.07	.27	-.14
Interview Composite[a]	.15	--	.22	--	.11	--	.25	--
ExAct General	.11	.18	.17	.27	.06	.18	.12	.38
ExAct Computer	.07	.02	.11	.04	.11	-.11	.23	-.24
PFF PPW Achievement	.26*	.10	.38	.15	.13	.09	.27	.21
PFF PPW Civilian Education	.23*	-.03	.33	-.04	.20*	-.12	.41	-.26
PFF Disciplinary Actions	-.20*	-.30*	-.29	-.42	-.17*	-.17	-.37	-.36
PFF APFT score (weighted)	.18*	.28*	.27	.40	.24*	.20	.49	.41
PFF APFT score	.13	.10	.20	.15	.14	.04	.30	.09
PFF Awards (weighted)	.08	.10	.12	.15	-.03	.00	-.06	.01
PFF Achievement Certificates	.17	.18	.25	.27	.09	.22	.19	.45
PFF Military Training	.14	.22	.21	.33	.16	.19	.33	.39
PFF PPW Military Education	.10	.01	.15	.02	.11	-.06	.24	-.14
PFF Awards	.09	.20	.13	.29	-.01	.06	-.03	.13
PFF Weapons Qualification	-.03	.11	-.05	.17	.00	.08	.00	.17
PFF Memoranda/Letters	-.01	-.04	-.02	-.07	.01	.05	.03	.10
AFQT Score	.04	.02	.06	.03	.13	-.04	.28	-.09
GT Score	.12	.03	.18	.05	.11	.03	.23	.07
AIM Work Orientation	.34*	--	.48	--	.32*	--	.60	--
AIM Leadership	.26*	--	.38	--	.28*	--	.55	--
AIM Physical Conditioning	.21*	--	.31	--	.14	--	.29	--
AIM Adjustment	.14	--	.21	--	.12	--	.27	--
AIM Agreeableness	.13	--	.20	--	.12	--	.27	--
AIM Dependability	.06	--	.10	--	.02	--	.05	--
BIQ Emergent Leadership	.21*	--	.31	--	.28*	--	.55	--
BIQ Hostility to Authority	-.16	--	-.24	--	-.14	--	-.30	--
BIQ Openness	-.09	--	-.13	--	.00	--	.01	--
BIQ Manipulativeness	.04	--	.06	--	-.01	--	-.01	--
BIQ Tolerance for Ambiguity	.03	--	.05	--	.14	--	.29	--
BIQ Social Perceptiveness	.03	--	.05	--	.13	--	.27	--
BIQ Social Maturity	.01	--	.02	--	.01	--	.02	--
BIQ Interpersonal Skill	.02	--	.02	--	.06	--	.13	--

Note. Dashes indicate the predictor measure was not administered to E6 soldiers. n_{E5} = 66-137, n_{E6} = 36-74.

[a]Interview composite does not include the MOS/Occupation-Specific Knowledge and Skill score to maximize sample size.

[b]Correlations are corrected for unreliability in the criterion (see text for details). They are based on the single rater reliabilities for soldiers who were each rated by at least two supervisors, r_{IRR1} = .34 for the Observed Performance Rating Scales composite and r_{IRR1} = .16 for the Expected Future Performance Rating Scales composite.

*$p < .05$.

One salient finding was that correlations of some predictors with expected future performance were opposite from one grade to the next, such that certain measures showed moderate positive correlations for E5 soldiers and moderate negative correlations for E6 soldiers (i.e., SJT Form B, ExAct Computer, PFF Military Education, PFF Civilian Education, and AFQT). This may suggest an underlying factor affected the way supervisors rated E5 and E6 soldiers on expected future performance.

Situational Judgment Test

A meta-analysis study by McDaniel et al. (2001) found a correlation of .36 (corrected for measurement error in the criteria) between SJT scores and performance measures (93% employed supervisory ratings or rankings, and the remainder used production data as performance measures). The corrected correlations for the SJT (Form A) presented in Table 7.6 show comparable relationships. However, none of the uncorrected correlations between the SJT Forms A and B and observed or future performance was significant.

Semi-Structured Interview

The corrected correlations between the interview and the composite scores for observed performance (r = .22) and expected future performance (r = .25) were in the low to moderate range. This is fairly consistent with the pattern of relationships found in the ECQUIP project. Specifically, in ECQUIP, the uncorrected correlation between the ECQUIP interview and supervisory Behaviorally Anchored Rating Scales (BARS) scores (both served as criterion measures) was .16, very similar to the uncorrected correlations found in the NCO21 data (r = .15 Observed, r = .11 Future). Thus, it is not surprising that strong relationships were not found between the interview scores and the criteria for the field test data because the low to moderate relationships could be due, in part, to differences in the types of performance being rated. In particular, the supervisory ratings of observed performance were based on the soldier's typical performance (i.e., what the soldier "has done") whereas the interviewers' ratings were based on examples of the soldier's self-reported performance, which were usually examples of "best" performance (or what the soldier "could do" in a situation).

Experiences and Activities Record

In Project A, self-report instruments assessing biodata (e.g., ABLE) and archival information (e.g., Personnel File Form) provided information relevant for predicting soldier performance. Thus, it would be expected that the ExAct and PFF21 scores would be correlated with the supervisory rating composites in this research. Indeed, the ExAct General scores show low positive relationships with the criteria for E5 soldiers (r = .17 Corrected Observed, r = .12 Corrected Future) and stronger positive relationships with the criteria for E6 soldiers (r =.27 Corrected Observed, r =.38 Corrected Future). These results provide some support for the criterion-related validity for the ExAct General score. The results also support previous research that found biodata instruments to predict similar performance-related criteria in the military (e.g., effectiveness ratings; see Trent & Laurence, 1993, for a review). The corrected correlations between the ExAct Computer score and the Observed Performance Rating Scales composite were low or negative across grades. The relatively large negative correlation (-.24) at the E6

level is counterintuitive. Perhaps supervisors feel that soldiers at the E6 level should focus more on leadership than computer skills, even in the future Army.

Personnel File Form 21

The corrected correlations between the PFF21 and composite rating scale scores show that several PFF21 scores have moderate to strong relationships with one or more of the criteria for E5 and/or E6 soldiers. In particular, the weighted APFT score shows consistently strong correlations across grades for both performance rating composite scores.

Previous research from the ECQUIP project found low negative (uncorrected) correlations between Disciplinary Actions and BARS rating scores; thus, similar results were expected in the present research. As expected, we found significant moderate to strong negative corrected correlations between Disciplinary Actions and the criterion composite scores across grades. The correlation coefficients for Awards (weighted), Weapons Qualification, and Memoranda/Letters, however, were all generally low (i.e., $r = .17$ or below). Further, the Awards (weighted) score did not appear to have as strong of a relationship to the criteria as did the raw score. The utility of these particular scales will be more apparent after additional data are collected during the validation effort.

ASVAB

Although the correlation coefficients for AFQT and GT showed little relationship with the Observed Performance Rating Scale composite score for E5 or E6 soldiers, the correlations were positive and not dramatically inconsistent with the results found in the ECQUIP project. Specifically, the corrected correlation between AFQT and the observed performance composite for E5 soldiers ($r = .06$) was similar to the uncorrected correlation found between AFQT and the ECQUIP Supervisory BARS ($r = .07$) (Peterson et al., 1997) and between AFQT and the supervisory/peer ratings collected in Project A ($r = .11$) (J. Campbell & Knapp, 2001). The corrected correlation with GT was even higher at the E5 level ($r = .18$).

The corrected correlation between AFQT and the Expected Future Performance Ratings composite score in the present effort showed a low to moderate relationship for E5 soldiers ($r = .28$), but not for E6s ($r = -.09$). The GT correlations were .23 for E5 and .07 for E6 soldiers. However, none of the uncorrected correlations between AFQT or GT and the observed or future performance scores were significant for E5 or E6 soldiers. Overall, restriction of range in scores does not appear to be the source of the low correlations with the criterion measures, as these scores showed a fair amount of variability within grades. Perhaps part of the reason for the low correlations is ASVAB's ability to predict "can-do" types of performance and the rating scales assessing "have-done" types of performance, which could be influenced by factors such as motivation.

Assessment of Individual Motivation (AIM)

Nearly all corrected bivariate correlations for the AIM were moderate to high for both performance composites. In particular, the Work Orientation and Leadership scales showed high correlations for both instruments (ranging from $r = .38$ to $r = .60$). Dependability scale scores

demonstrated a low relationship with both criterion scores. This pattern of results is consistent with that found in ECQUIP for uncorrected correlations between ABLE (i.e., the predecessor to AIM) and Supervisory BARS scores.

Biographical Information Questionnaire (BIQ)

There was a significant uncorrected correlation between the BIQ Emergent Leadership scale and both performance criteria. When these correlations are corrected, the relationship is particularly strong for the Expected Future Performance Rating scale composite. In addition, the results showed moderate negative corrected correlations between BIQ Hostility to Authority and both performance criterion measures. Two other BIQ scales (Tolerance for Ambiguity, Social Perceptiveness) were moderately associated (corrected correlation) with expected future performance, but showed little to no relationship with the ratings of observed performance. The correlation with Tolerance for Ambiguity makes sense because one of the future performance scenarios focuses on the ability to adapt to changing conditions or conditions for which there is little information available. However, it is unclear why this pattern of correlations exists with Social Perceptiveness. Although it is merely speculation, perhaps supervisors are basing their expected future performance ratings, in part, on personality factors (e.g., Social Perceptiveness, Hostility to Authority), and they are doing this to a lesser extent in their ratings of observed performance.

Summary

As a whole, the results presented in this chapter provide preliminary evidence of the relationships among the predictors and between the predictors and criteria. For the most part, measures assessing similar constructs correlated with each other, and most predictor measures had at least one scale that correlated with at least one of the criterion scores. Personality and experience variables had larger correlations with the criterion measures than did *g*-loaded (i.e., targeting cognitive ability) predictors such as AFQT or the SJT. Perhaps this is due to an Army value system that emphasizes motivation, integrity, and physical fitness.

Sample sizes with complete data on all instruments were insufficient to conduct the types of construct validity and criterion-related validity analyses planned for the validation data collection (e.g., evidence of convergent and divergent validity, multiple regressions, examinations of incremental validity of predictors). This preliminary analysis of interrelationships will provide useful information for the development of hypotheses in the validation research.

CHAPTER 8: SUMMARY

In this report, we have described the selection and development of instruments to be used as predictor and criterion measures in the NCO21 criterion-related validation research effort. The predictor measures are designed to be suitable for incorporation into the Army NCO semi-centralized promotion system.

Predictor Measures

The project team identified seven predictor measures for use in the NCO21 project. Three measures (ASVAB, AIM, and BIQ) are operational tests used in the Army for other purposes. Experimental versions of the AIM and BIQ were prepared for use in the present research. Four measures—SJT (and its close cousin, SJT-X), ExAct, PFF21, and a semi-structured interview—were developed for this project. Most of these instruments, however, made use of relevant, previously-developed materials and items.

Table 8.1 shows the seven predictor measures and indicates which of the 38 NCO21 KSAs are assessed by each. A checkmark indicates that the KSA is explicitly targeted by the instrument. An "X" indicates we would expect scores on the measure to correlate with direct measures of the KSA, even though the KSA is not explicitly targeted.

Only three KSAs have no coverage, either directly or indirectly. These are either low priority KSAs as identified by the Phase II expert panels (e.g., Safety Consciousness) or ones that would require very different measurement strategies than those that were adopted (e.g., Psychomotor Aptitude). A number of the higher priority KSAs, however, are addressed by several of the predictor measures.

Criterion Measures

The Observed Performance Rating Scales cover all 27 of the NCO21 performance requirements. (Recall the 27 performance requirements are a subset of the 38 KSAs.) The 27 performance requirements, however, were consolidated into a more manageable set of 19 areas to be rated.

The Expected Future Performance Rating Scales are not intended to measure the specific performance requirements, per se. Rather, they ask for evaluations of overall performance, given specific sets of alternative conditions.

Under a separate contract effort, Aptima researchers are developing a computer-based simulation that may also be used as a criterion measure for some of the validation research participants. As of this writing, the simulation is in the fairly early stages of development, so the final set of performance requirements that will be assessed is uncertain. A major goal of the developers, however, is to assess at least two futuristic performance requirements that are probably not captured well with supervisor ratings of current performance (i.e., Knowledge of the Inter-Relatedness of Units, Management/Coordination of Multiple Battlefield Functions).

Table 8.1. Measurement Methods by KSAs

KSA	Measurement Method						
	ASVAB	SJT	ExAct	PFF21	Interview	AIM	BIQ
General Cognitive Aptitude	✓	X			X		
Working Memory	X						
Basic Math Facility	✓						
Basic Electronics Knowledge	✓						
Basic Mechanical Knowledge	✓						
Spatial Relations Aptitude	✓[a]						
Perceptual Speed & Accuracy	✓[a]						
Psychomotor Aptitude							
Problem-Solving/Decision Making	X	✓					
Information Management	X						
Writing Skill	X		✓	X			
Oral Communication Skill					✓		
MOS-Specific Knowledge & Skill	X			X	✓		
Common Task Knowledge & Skill	X			X			
Safety Consciousness							
Computer Skills			✓				
Knowledge of the Inter-Relatedness of Units	X	SJT-X					
Management and Coordination of Multiple Battlefield Functions	X						
Motivating, Leading, and Supporting Individual Subordinates	X	✓	✓		✓[b]	X	
Directing, Monitoring, and Supervising Individual Subordinates	X	✓	✓			X	
Training Others	X	✓	✓		✓[b]	X	
Modeling Effective Performance		X	X	X		X	
Relating to and Supporting Peers		✓			✓		
Team Leadership		✓	✓		✓[b]		
Concern for Soldier Quality of Life		✓					
Cultural Tolerance		✓					
Selfless Service Orientation							
Level of Effort and Initiative on the Job			✓	X	✓	X	
Need for Achievement						✓	
Conscientiousness/Dependability						✓	✓
Adherence to Regulations, Policies, and Procedures				✓		X	X
Level of Integrity and Discipline on the Job				✓	✓	X	X
Emotional Stability						✓	
Adaptability					✓	✓	

Table 8.1. Measurement Methods by KSAs (Continued)

KSA	Measurement Method						
	ASVAB	SJT	ExAct	PFF21	Interview	AIM	BIQ
General Self-Management Skill					✓		
Self-Directed Learning Skill				X	✓		
Physical Fitness				✓		X	
Military Presence					✓		

Note. 3 = designed to measure; X = expected to correlate.

[a]Spatial relations and perceptual speed and accuracy are measured by the Assembling Objects subtest which is now included as an experimental test on the CAT-ASVAB.

[b]Several KSAs were combined for measurement via the interview.

Validation Data Collection Plans

Plans are to collect validation data from seven Army installations from April through August, 2001. The goal is to collect complete predictor data for E4 soldiers, complete predictor and criterion data for E5 soldiers, and partial predictor data (all except the interview) and complete criterion data for E6 soldiers.

Troop Support Requests

A total of 2,455 soldiers, along with two supervisors for each of the E5 and E6 soldiers, have been requested to participate. Table 8.2 summarizes the requests for E4-E6 soldiers at each of the seven sites. In addition to the E4-E6 soldiers and their supervisors, participating Army installations have also been asked to provide 10 senior NCOs to participate as interviewers.

Table 8.2. NCO21 Validation Data collection Troop Support Request Summary

Installation	Dates	E4 request	E5 request	E6 request	Total request
Ft. Hood, TX	2-6 Apr	90	180	135	405
Ft. Bragg, NC	9-13 Apr	90	180	135	405
Ft. Lewis, WA	23-27 Apr	90	170	125	385
Ft. Riley, KS	21-24 May	90	135	90	315
Ft. Campbell, KY	4-8 Jun	90	135	90	315
Ft. Carson, CO	18-22 Jun	90	135	90	315
Ft. Stewart, GA	20-24 Aug	90	135	90	315
Total		630	1,070	755	2,455

Current plans call for administering the Aptima computerized simulation criterion measure to 30 soldiers at Fort Stewart. These soldiers will have also participated in the NCO21 data collection during the same time period. Aptima researchers may also try to collect additional simulation data at a later time, along with at least a subset of the NCO21 predictor measures.

Overview of On-Site Data Collection Activities

Separate 3-hour written test sessions have been scheduled for E4, E5, and E6 soldier participants. Supervisors of the E5 and E6 soldiers will report to another classroom or area to provide performance ratings. In separately scheduled individual 45-minute sessions, E4 and E5 soldiers will be given the semi-structured interview. A sample data collection schedule is shown in Figure 8.1.

The E4/E5/E6 soldier and supervisor sessions will involve the same initial steps. The data collection team will introduce themselves, give a brief project briefing, read a Privacy Act statement, and have participants complete a short Background Information Form.

		Room 1	Room 2	Room 3
Day 1	am	45 E6s	Supervisors	Interviewer training
	pm	45 E6s	Supervisors	Interviews
Day 2	am	45 E5s	Supervisors	Interviews
	pm	45 E5s	Supervisors	Interviews
Day 3	am	45 E5s	Supervisors	Interviews
	pm	45 E4s	Supervisors*	Interviews
Day 4	am	45 E4s	Supervisors*	Interviews
	pm		Supervisors*	Interviews
Day 5	am		Supervisors*	Interviews
			Supervisors*	Interviews

*These are supervisor make-up sessions.

Figure 8.1. Sample validation data collection schedule.

A list of instruments to be given in these sessions is provided in Table 8.3.With one exception, the E4-E6 soldiers will get the same forms in the 3-hour test session. The exception is that only the E6 participants will get the SJT-X (in addition to the SJT).

Table 8.3. Instruments Administered in Soldier Test Sessions

- Background Information Form
- Experiences & Activities Record (ExAct)
- Personnel File Form-21 (PFF21)
- Situational Judgment Test (SJT)
- SJT-X *(E6 soldiers only)*
- Assessment of Individual Motivation (AIM)
- Biographical Information Questionnaire (BIQ)

The interviews will be administered by a group of 10 trained senior NCOs. The NCOs will be paired into two-person interview teams, so five soldiers can be interviewed at any given time.

A test administrator's manual has been developed, and data collection staff will have participated in a 4-6 hour training program prior to collecting project data. This training provides instructions for preparing materials, scripts for administering the various measures, and instructions for handling the data and instruments.

Analysis and Final Recommendations

Whereas the field test data analyses focused on finalization of the project instruments, the validity sample analyses will focus on the quality of the final measures in terms of their psychometric properties and relationships among other measures. In particular, we will attempt to demonstrate the criterion-related validity of the various experimental predictors. These results will be contrasted with the estimated validity of the currently used promotion criteria (to the extent that this can be modeled using data from the PFF21).

Table 8.4 summarizes some of the major research questions we will address in the validation data analysis effort.

Table 8.4. Summary of Major Research Questions

- What is the psychometric quality of the predictor and criterion measures?
- What are the relationships among the measures within each domain?
- What are the major dimensions of performance?
- To what extent does performance on the predictors relate to performance on various aspects of the job?
- What combination of predictors best predicts job performance?
- How does the best combination of predictors compare to the current set of predictors?

Final recommendations to the Army will be based not only on the results of the validation effort data analyses, but also on feedback from research participants, Army sponsors, and expert panelists; ideas generated during the course of the research; and practical considerations regarding the ease with which various measures could be implemented.

REFERENCES

Borman, W. C., Motowidlo, S. J., Rose, S. R., & Hanser, L. M. (1985). *Development of a model of soldier effectiveness* (ARI Technical Report 741). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Campbell, C. H., Ford, P., Rumsey, M. G., Pulakos, E. D., Borman, W. C., Felker, D. B., De Vera, M. V., & Riegelhaupt, B. J. (1990). Development of multiple job performance measures in a representative sample of jobs. *Personnel Psychology, 43*, 277-300.

Campbell, J. P. (Ed.) (1987). *Improving the selection, classification, and utilization of Army enlisted personnel: Annual report, 1985 fiscal year* (ARI Technical Report 746). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Campbell, J. P., & Knapp, D. J. (2001). *Exploring the Limits in Personnel Selection and Classification.* Mahwah, NJ: Lawrence Erlbaum Associates.

Carlson, K. D., Scullen, S. E., Schmidt, F. L., Rothstein, H., & Erwin, F. (1999). Generalizable biographical data validity can be achieved without multi-organizational development and keying. *Personnel Psychology, 52*, 731-755.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98-104.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Ft. Worth, TX: Harcourt Brace Jovanovich College Publishers.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Ford, L., Campbell, R., Campbell, J. P., Knapp, D., & Walker, C. (2000). *21st-century soldiers and noncommissioned officers: Critical predictors of performance* (ARI Technical Report 1102). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Hansen, C. P. (1989). A causal model of the relationship among accidents, biodata, personality, and cognitive factors. *Journal of Applied Psychology, 74*, 81-90.

Hedlund, J., Williams, W.M., Horvath, J.A., Forsythe, G.B., Snook, S., Wattendorf, J., McNally, J.A., Sweeney, P.J., Bullis, R.C., Dennis, M., & Sternberg, R.J. (1999). *Tacit knowledge for military leaders: Platoon Leader Questionnaire* (ARI Research Product 99-07). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Hough, L. E. (1984). Development and evaluation of the "Accomplishment Record" method of selecting and promoting professionals. *Journal of Applied Psychology, 59*, 135-146.

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology, 75*, 581-595.

Kilcullen, R. N., Chen, G., Zazanis, M. M., & Carpenter, T., & Goodwin, G. (1999, April). *Adaptable performance in unstructured environments.* Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Kilcullen, R. N., Mael, F. A., Goodwin, G. F., & Zazanis, M. M. (1999). Predicting U.S. Army Special Forces Field Performance. *Journal of Human Performance in Extreme Environments, 4*, 53-63.

Kilcullen, R. N., White, L. A., Mumford, M. D., & Mack, H. (1995). Assessing the construct validity of rational biodata scales. *Military Psychology, 7*, 17-28.

Kilcullen, R. N., White, L. A., Zacarro, S., & Parker, C. (2000, April). *Predicting managerial and executive performance.* Paper presented at the 15th annual meeting of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Laurence, J. H. (1990). ASP—what you can do for your country: Biodata and military selection. *Forensic Reports, 3*, 169-178.

Mael, F. A. (1991). A conceptual rationale for the domain and attributes of biodata items. *Personnel Psychology, 44*, 763-792.

Mael, F. A. (1994). If past behavior really predicts future, so should biodata's. In M. G. Rumsey and C. B. Walker (Eds.), *Personnel selection and classification.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Mael, F. A., & Ashforth, B. E. (1995). Loyal from day one: Biodata, organizational identification, and turnover among newcomers. *Personnel Psychology, 48*, 309-333.

Mael, F. A., & Hirsch, A. C. (1993). Rainforest empiricism and quasi-rationality: Two approaches to objective biodata. *Personnel Psychology, 46*, 719-738.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86,* 730-740.

McManus, M. A., & Kelly, M. L. (1999). Personality measures and biodata: Evidence regarding their incremental predictive value in the life insurance industry. *Personnel Psychology, 52*, 137-148.

Mitchell, T. W. (1994). The utility of biodata. In G. S. Stokes & M. D. Mumford (Eds.), *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (pp. 485-516). Palo Alto, CA: Consulting Psychologists Press.

Mumford, M. D., & Owens, W. A. (1987). Methodology review: Principles, procedures, and findings in the application of background measures. *Applied Psychological Measurement, 2*, 1-31.

Peterson, N. G., Anderson, L. E., Crafts, J. L., Smith, D. A., Reynolds, D. H., Motowidlo, S. J., Rosse, R. L., Dela Rosa, M. R., & Waugh, G. W. (1997). *Expanding the concept of quality in personnel: Final report.* Washington, DC: American Institutes for Research.

Peterson, N. G., Anderson, L. E., Crafts, J. L., Smith, D. A., Motowidlo, S. J., Rosse, R. L., Waugh, G.W., McCloy, R., Reynolds, D. H., & Dela Rosa, M. R. (1999). *Expanding the concept of quality in personnel: Final report* (ARI Research Note 99-31). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Quartetti, D.A., & Tsacoumis, S. (2000). *Social Security Administration Leadership Development Program structured interview* (FR-00-16). Alexandria, VA: Human Resources Research Organization.

Rumsey, M. G., Peterson, N. G., Oppler, S. H., & Campbell, J. P. (1996). What's happened since Project A: The future career force. *Journal of the Washington Academy of Sciences, 84,* 94-110.

Russell, T. L., Crafts, J. L., Peterson, N.G., Rohrback, M. R., Nee, M. T., & Mael, F. (1995). *Development of a roadmap for Special Forces selection and classification research* (ARI Technical Report 1033). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Russell, T. L., Crafts, J. L., Tagliareni, F. A., McCloy, R. A., & Barkley, P. (1996). *Job analysis of Special Forces jobs* (ARI Research Note 96-76). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428.

Stokes, G. S., & Toth, C. S. (1996). Background data for personnel selection. In R. S. Barrett et al. (Eds.), *Fair employment strategies in human resource management* (pp. 171-179). Westport, CT: Greenwood.

Trent, T., & Laurence, J.H. (1993). *Adaptability screening for the Armed Forces.* Washington, DC: Office of Assistant Secretary of Defense (Force Management and Personnel).

Vinchur, A. J., Schippmann, J. S., Switzer, F. S., & Roth, P. L. (1998). A meta-analytic review of predictors of job performance for salespeople. *Journal of Applied Psychology, 83,* 586-597.

White, L. A., & Young, M. C. (1998, August). *Development and validation of the Assessment of Individual Motivation (AIM).* Paper presented at the annual meeting of the American Psychological Association, Washington, DC

White, L. A., & Young, M. C. (2001, April). *Validation of a faking-resistance measure of temperament constructs.* Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, San Diego, CA.

White, L. A., Nord, R. D., Mael, F. A., & Young, M. C. (1993). The Assessment of Background and Life Experiences (ABLE). In T. Trent and J. H. Laurence (Eds.), *Adaptability screening for the armed forces*. Washington, DC: Office of Assistant Secretary of Defense (Force Management and Personnel).

White, L. A., Young, M. C., & Rumsey, M. G. (2001). Assessment of Background and Life Experiences (ABLE) implementation issues and related research. In J. P. Campbell and D. J. Knapp (Eds.), *Exploring the Limits in Personnel Selection and Classification.* Hillsdale, NJ: Erlbaum.

Young, M. C., Heggestad, E. D., Rumsey, M. G., & White, L. A. (2000, August). *Army Pre-implementation Research Findings on the Assessment of Individual Motivation (AIM).* Paper presented at the annual meeting of the American Psychological Association, Washington, DC.

Zickar, M. J., & Robie, C. (1999). Modeling faking good on personality items: An item-level analysis. *Journal of Applied Psychology, 84*, 551-563.

Appendix A
Observed Performance Rating Scales

Section I: Observed Performance Rating Scales

1. MOS/Occupation-Specific Knowledge and Skill						
How effectively does this soldier display job-specific knowledge and skill?						
Does not display the knowledge or skill required to perform many work assignments or tasks; is unaware of recent developments relevant to his/her MOS.		Displays adequate knowledge of most aspects of the job; has sufficient skills to handle moderately difficult problems and to get most assignments done properly; attempts to keep informed of most important developments in his/her MOS.			Is highly competent in performing the technical tasks for which he/she is responsible; has skills and technical knowledge necessary to handle difficult problems; strives to stay informed of latest developments in his/her MOS.	
LOW		**MODERATE**			**HIGH**	
1	**2**	**3**	**4**	**5**	**6**	**7**

2. Common Task Knowledge and Skill						
How effectively does this soldier display the necessary knowledge and skill to perform common tasks?						
Does not display the knowledge or skill required to perform common assignments or tasks (e.g., land navigation, field survival techniques, NBC protection).		Displays good knowledge of most common areas; has sufficient skills to handle moderately difficult problems and to perform common tasks properly.			Is highly competent in performing common tasks; possesses skills and knowledge necessary to handle most common tasks, even under difficult conditions.	
LOW		**MODERATE**			**HIGH**	
1	**2**	**3**	**4**	**5**	**6**	**7**

3. Computer Skills						
To what extent does this soldier display an understanding of computer systems, operating systems, and applications?						
Does not display any understanding of computers above basic usage or Windows-based applications; cannot troubleshoot even the most basic application errors.		Displays basic understanding of some operating systems (e.g., DOS, Windows NT); can troubleshoot basic application errors; can troubleshoot simple systems errors; understands computer terminology.			Is highly competent administrating most operating systems (e.g., DOS, Windows NT, Army specific); can troubleshoot serious application errors; can set up and troubleshoot computer systems; well versed in computer terminology.	
LOW		**MODERATE**			**HIGH**	
1	**2**	**3**	**4**	**5**	**6**	**7**

4. Writing Skill						
How effectively does this soldier prepare written materials?						
Usually writes in an awkward or confusing manner; uses incorrect grammar, punctuation, and spelling; often includes irrelevant information in the material; written products often require a lot of editing.		Typically writes logically but will occasionally make grammatical, punctuation, or spelling errors; usually includes most relevant information and tries to tailor the work to the audience; written products sometimes require editing.			Usually writes concisely, clearly, and logically; focuses on relevant issues; uses correct grammar, punctuation, and spelling; effectively tailors the work to the audience; written products require little or no editing.	
LOW		**MODERATE**			**HIGH**	
1	**2**	**3**	**4**	**5**	**6**	**7**

5. Oral Communication Skill						
How effectively does this soldier orally communicate?						
Speaks in an awkward or confusing manner; does not present ideas clearly; often rambles or strays to irrelevant topics; mispronounces words or terms; speaks too fast or too slow.		Usually expresses him or herself clearly and logically; makes few grammatical errors; typically gets information across effectively; generally speaks at an appropriate, smooth pace.			Always expresses him or herself clearly and logically; gets to the point quickly; uses correct grammar; appropriately tailors the presentation to the audience; focuses on relevant and important issues; always speaks fluently and at a smooth pace.	
LOW		**MODERATE**			**HIGH**	
1	**2**	**3**	**4**	**5**	**6**	**7**

6. Level of Effort and Initiative on the Job						
To what extent does this soldier put forth effort and initiative on the job/mission/assignment?						
Shows little effort or initiative to accomplish tasks; completes assignments carelessly; often fails to meet deadlines; rarely seeks out additional responsibilities or challenging tasks.		Demonstrates sufficient effort on most tasks and assignments; is usually reliable about completing assignments on time; puts forth extra effort when necessary; sometimes seeks out additional responsibilities, training, or challenging tasks.			Shows a lot of initiative and often puts forth extra effort to get tasks done effectively, even under difficult conditions; reliably accomplishes work on time; enthusiastically takes on challenging assignments and additional responsibilities.	
LOW		**MODERATE**			**HIGH**	
1	**2**	**3**	**4**	**5**	**6**	**7**

7. Adaptability		
How effectively does this soldier adapt to varying environments by modifying behavior, plans, or goals?		
Has difficulty functioning effectively in new situations; does not adapt quickly to new environments, people, or equipment; is easily frustrated in situations that do not go as planned.	Is able to function adequately in new situations; modifies behavior when faced with unexpected events or conditions; adapts fairly readily to new people, situations, or equipment.	Thinks and acts quickly in response to changes in the environment; often develops innovative and imaginative approaches to dealing with unexpected events; can effectively change plans when the situation requires it.
LOW	**MODERATE**	**HIGH**
1 2	**3 4 5**	**6 7**

8. Self-Management and Self-Directed Learning Skill		
How effectively does this soldier self-manage his/her job responsibilities, training and career development, and personal responsibilities?		
Makes little or no effort to balance work and personal responsibilities; uses finances irresponsibly; ignores or otherwise fails to participate in relevant career training opportunities; needs constant supervision; fails to seek advice when needed.	Shows effort to manage work and personal responsibilities; typically uses finances responsibly; participates in required courses/training; attempts to work on problem areas when encouraged to do so; can usually work independently; seeks advice when needed but sometimes from inappropriate sources.	Effectively manages work and personal responsibilities; demonstrates exceptional financial responsibility; studies and works hard during off-duty hours to improve job-related skills; actively seeks additional responsibilities to improve job skills and increase chance of promotion; works well without supervision; willingly seeks advice when appropriate.
LOW	**MODERATE**	**HIGH**
1 2	**3 4 5**	**6 7**

9. Demonstrated Integrity, Discipline, and Adherence to Army Procedures		
To what extent does this soldier adhere to Army procedures and values, and demonstrate integrity, ethical behavior, and self-discipline on the job?		
Is disrespectful toward superiors; is sometimes dishonest; has difficulty accepting and following superiors' orders; makes up excuses to avoid assignments; fails to take responsibility for his/her job-related errors; often fails to follow rules, policies, and regulations; takes unnecessary risks that endanger the safety of self and/or others.	Is usually respectful to superiors; is generally honest; obeys direct orders; takes responsibility for most job-related mistakes he/she makes; usually attempts to follow applicable rules, policies, and regulations; typically avoids unnecessary risks and notices potential safety hazards.	Is always respectful to superiors; is honest about work matters, even when it may go against personal interests; obeys orders; ensures others are not blamed for his/her mistakes; carefully follows rules, policies, and regulations; tries to make sure others follow the rules; takes steps to protect self and others from safety risks.
LOW	**MODERATE**	**HIGH**
1 2	**3 4 5**	**6 7**

10. Acting as a Role Model		
To what extent does this soldier set a good example for others to follow in terms of physical fitness, military bearing, and appropriate behavior?		
Is generally overweight or in poor physical condition; avoids exercise; often dresses sloppily; displays poor military bearing; sets a poor example for others to follow and fails to model even minimally acceptable behavior as a soldier.	Meets basic standards for physical fitness; dresses properly, maintaining Army standards; usually displays good military bearing; attempts to set a good example of soldier behavior for others to follow.	Exercises consistently to maintain excellent physical fitness; always dresses sharply in correct uniform; consistently maintains excellent military bearing; sets an outstanding example for others by exceeding the standards for appropriate military behavior.
LOW	**MODERATE**	**HIGH**
1 2	**3 4 5**	**6 7**

11. Relating to and Supporting Peers		
How effectively does this soldier relate to and support peers?		
Tends to be rude, selfish, and insensitive to peers' concerns; generally fails to provide assistance to others, even when there is a clear need to do so; may force his/her approach to tasks on others without seeking input.	Usually courteous and tactful when dealing with peers; provides assistance to others, especially when it is clear that help is needed; tries to develop approaches to tasks that take into account obvious differences of opinion.	Always treats peers in a courteous and tactful manner; offers assistance without waiting to be asked, even in situations that involve complicated interpersonal situations; actively seeks out peers' opinions and incorporates peers' ideas into own plans.
LOW	**MODERATE**	**HIGH**
1 2	**3 4 5**	**6 7**

12. Cultural Tolerance		
How effectively does this soldier demonstrate tolerance and understanding of other cultural and social backgrounds both in the context of the diversity of U.S. Army personnel and interactions with foreign nationals?		
Does not understand or show respect for other cultural practices or beliefs; makes insensitive comments or slurs to others based on social or cultural differences, (e.g., racial heritage, religious beliefs, ethnic customs, language); cannot work, socialize, or communicate effectively with others from different backgrounds.	Recognizes need to be tolerant and respectful of other cultural, ethnic, and belief systems but does not always demonstrate understanding of social and cultural diversity; willing to work, communicate, and perhaps socialize with others from different backgrounds but does not do so easily.	Shows tolerance, understanding, and respect for other cultural, ethnic, and belief systems; shows respect for social and cultural diversity, (e.g., racial heritage, religious beliefs, ethnic customs, language); easily works, socializes, and communicates well with others regardless of differences in background.
LOW	**MODERATE**	**HIGH**
1 2	**3 4 5**	**6 7**

13. Selfless Service Orientation		
To what extent does this soldier display a selfless service orientation?		
Fails to support team or group; has a "looking out for number one" attitude; explicitly asks for credit for unselfish behavior.	Supports team or group when called upon to do so, but usually waits until asked; puts group or team goals ahead of own goals when it is easy to do so.	Willingly commits to the greater good of the team; willingly puts group or team goals ahead of individual goals when appropriate; does not expect credit for unselfish behavior.
LOW	**MODERATE**	**HIGH**
1 2	**3 4 5**	**6 7**

14. Leadership Skills		
To what extent does this soldier demonstrate strong leadership skills by effectively motivating, supporting and supervising individuals and being an effective team leader?		
Fails to support subordinates; does not reward effective behavior or provide useful feedback to improve performance; assigns duties unfairly; rarely makes sure assignments are understood and completed; does not communicate team goals; fails to lead team to adapt to mission changes; fails to resolve conflicts or does so unfairly.	Usually supports subordinates and rewards effective behavior; provides feedback to improve performance, but it is not always helpful; generally assigns work fairly; typically makes sure subordinates' work meets standards; communicates team goals but not always clearly; leads team to adapt to mission changes but takes time/effort to do so; attempts to resolve conflicts fairly.	Always supports subordinates and rewards effective behavior; maintains high morale; provides helpful feedback to improve performance; always assigns work fairly; always makes sure subordinates' assignments are understood and completed; clearly communicates team goals; leads team to adapt quickly to mission changes; resolves conflicts among subordinates fairly.
LOW	**MODERATE**	**HIGH**
1 2	**3 4 5**	**6 7**

15. Concern for Soldier Quality of Life		
How effectively does this soldier show consideration for subordinates' quality of life?		
Generally ignores subordinates' personal needs, constraints, and values; ignores or is insensitive to potential conflicts between subordinates' personal needs and duty demands; fails to show concern for the well-being of subordinates' personal lives.	Usually is aware of and attempts to help resolve conflicts between subordinates' work and personal needs; is sometimes sensitive to potential work/personal conflicts and attempts to help subordinates avoid such situations; shows basic awareness of subordinates personal needs, constraints, and values.	Has keen awareness of subordinates personal needs, constraints, and values; takes extra steps to resolve and avoid subordinate work/personal life conflicts; shows genuine concern for the well-being of subordinates' personal lives.
LOW	**MODERATE**	**HIGH**
1 2	**3 4 5**	**6 7**

16. Training Others						
How effectively does this soldier provide relevant training experiences for subordinates?						
Is unaware of or ignores individual or unit training needs; fails to provide training experiences or gives subordinates inappropriate training; does not prepare well for formal training situations; fails to guide subordinates on technical training matters.		Usually ensures that important subordinate training needs are met when made aware of such needs; uses existing classroom or on-the-job training techniques; prepares as required for training sessions; sometimes guides and tutors subordinates on technical matters.			Actively seeks to be aware of individual or unit training needs; always makes time to provide relevant formal and informal training experiences for subordinates; prepares thoroughly for training sessions; effectively guides and tutors subordinates on technical matters.	
LOW		**MODERATE**			**HIGH**	
1	**2**	**3**	**4**	**5**	**6**	**7**

17. Coordination of Multiple Units and Battlefield Functions						
To what extent does this soldier demonstrate knowledge of the interrelatedness among different units (including his/her own unit), as well as how to coordinate multiple battlefield functions?						
Cannot apply or coordinate multiple battlefield functions such as direct/indirect fires, communications, intelligence, and combat service support (CSS) to achieve tactical goals; shows little or no ability to understand how one unit's actions can affect the performance of other units; does not see how his/her unit's operations relate to the overall system.		Can apply and coordinate multiple battlefield functions (e.g., direct/indirect fires, communications, intelligence, CSS) with assistance; usually recognizes how one unit's actions can affect the performance of other units; understands how some goals and operations of own unit and other units relate but has difficulty analyzing the overall system.			Can independently apply and coordinate multiple battlefield functions (e.g., direct/indirect fires, communications, intelligence, and CSS) to achieve tactical goals; clearly understands how one unit's actions can affect the performance of other units; can quickly and accurately analyze how goals and operations of own unit relate to the overall system.	
LOW		**MODERATE**			**HIGH**	
1	**2**	**3**	**4**	**5**	**6**	**7**

18. Problem-Solving/Decision Making Skill						
How effectively does this soldier react to new problem situations and make reasonable, informed decisions regarding solutions?						
Usually reacts to new problem situations with frustration and confusion; fails to apply previous experience and training or realize their relevance; blindly applies rules or strategies without regard to the uniqueness of the situation; fails to assess costs or benefits of alternative solutions before making decisions.		Often reacts to new problem situations by applying previous experience or education/training, but does not always do so effectively; seldom applies rules or strategies blindly; attempts to assess costs and benefits of alternative solutions but does not always make timely decisions; has trouble making appropriate decisions with incomplete information.			Consistently reacts to new problem situations by applying previous experience and previous education/training appropriately and effectively; does not apply rules or strategies blindly; assesses costs and benefits of alternative solutions and makes timely decisions even with incomplete information.	
LOW		**MODERATE**			**HIGH**	
1	**2**	**3**	**4**	**5**	**6**	**7**

19. Information Management		
How effectively does this soldier monitor, interpret, and redistribute information received from multiple sources (especially in a digitized environment)?		
Easily experiences information overload; has trouble monitoring and interpreting multiple information sources; is unable to cope with a digitized environment; is inefficient or unable to process information and prepare it for redistribution so that it is useable by others.	Usually can handle a fair amount of information effectively; often able to effectively monitor multiple information sources, but can become overwhelmed by the speed of communication provided by digitized equipment; is able to process information and redistribute it for use by others, but fails to effectively combine or exclude information.	Can monitor, interpret, and redistribute large amounts of information received from multiple sources, especially in digitized environments; processes information effectively so that it is optimally useful to others; does not readily experience information overload.
LOW	**MODERATE**	**HIGH**
1 2	**3 4 5**	**6 7**

Section II: Overall Effectiveness

Please read the description below of overall soldier effectiveness and then rate how effective each soldier is by marking the appropriate number.

Overall Effectiveness How effectively does this soldier perform overall?		
Performs poorly in important effectiveness areas; does not meet standards for soldier performance compared to peers at same experience level.	Performs adequately in important effectiveness areas; meets standards and expectations for soldier performance compared to peers at same experience level.	Performs excellently in all or almost all effectiveness areas; exceeds standards and expectations for soldier performance compared to peers at same experience level.

Section III: Senior NCO Potential

On this rating, evaluate each soldier on his or her potential effectiveness as a senior NCO (E-7 to E-9). At this point, you are not to rate on the basis of present performance and effectiveness, but instead, indicate how well each soldier is likely to perm as a senior NCO in his or her MOS (assume each will have an opportunity to be a senior NCO). Thus, the "overall effectiveness" rating you completed in Section II and this rating of senior NCO potential may not necessarily agree closely.

Senior NCO Potential Which of the following best describes each soldier's senior NCO potential?		
Would likely be a bottom-level performer as a senior NCO.	Would likely be an adequate performer as a senior NCO.	Would likely be a top-level performer as a senior NCO.

Appendix B

Expected Future Performance Rating Scales

Expected Performance Under Future Army Conditions

Instructions

In this booklet, you will read several scenarios that describe some of the major changes predicted to occur in the future Army. After you read each scenario please rate how effectively you would expect each soldier to meet those future NCO requirements. Note that actual future Army conditions may differ from these scenarios.

Use the separately provided scannable sheet to record your ratings.

Scenario #1: Increased Requirements for Self-Direction and Self-Management

The predicted changes in missions, technology, structure, and tactics will require that NCOs have a greater ability to guide their own professional development and manage their personal affairs (e.g., family concerns and financial matters). Obviously, increasing mission diversity and frequency will be disruptive. For example, frequent deployments away from U.S. home bases will require a strong ability to manage personal matters effectively. In addition, the restructuring of the Army into smaller, more independent units will require that NCOs have a greater ability to take initiative in their actions and make their own decisions without direct supervision. Finally, due to greater technological change and more frequent changes in missions, there is an expectation that individual NCOs will need to assume more and more responsibility for their own training. That is, they will be required to identify their own training needs and to seek out training experiences that meet these needs. They will need to evaluate their own training accomplishments and take corrective steps if necessary.

1. How effectively would you expect the soldier to meet these future NCO requirements?

Not likely to meet the NCO demands described under these conditions.		Likely to be generally successful, but will struggle to meet the NCO demands described under these conditions.			Likely to successfully meet or exceed NCO demands described under these conditions.	
LOW		**MODERATE**			**HIGH**	
1	**2**	**3**	**4**	**5**	**6**	**7**

Scenario #2: Use of Computers, Computerized Equipment, and Digitized Operations

The digitization of the Army that started in the mid-1990s will increase and become more widespread by 2010. Commercial applications of personal computers (PCs), laptops, and small hand-held devices will become the standard means for communicating and relaying information for all soldiers, in all jobs, at all levels. Specialized military applications of computers will become more widespread and will be found on all tactical vehicles and weapons systems. Voice recognition will provide essentially hands-free operation for crewmembers. Individualized applications, available to dismounted soldiers in a variety of roles, will provide automated links for information flow in tactical settings. In addition, a tactical Internet will make it possible for operators to link to each other at all levels and locations in real time. Automation will have a serious impact on the logistical and service support functions of the Army in that most aspects of supply, maintenance, and transport will use some form of computerized system. These will start with the user of the service or supply and be linked upwards to the depot level and beyond.

While much of the focus will be on computer hardware, the truly significant advancements in technology will involve the development of specialized software. These programs will cover a variety of functions such as land navigation, orders preparation, after action analysis, and information sorting and processing. This specialized software could change how soldiers function at all levels. The Army will likely be able to automate many of the current manual functions, giving greater skills and abilities to more individuals. At the same time, specialized software will require specialized input and manipulation.

Computerization and automation will not be foolproof. System failures, clutter, jamming, hacking, interceptions, and false information are all risks that come with the use of computer-based communications. The need for back-up manual knowledge, alternate procedures, fail-safe checks, and trouble-shooting skills will place increased demands on soldier knowledge and performance. NCOs and officers will need to be able to oversee and monitor systems used by lower-level operators and implementers. In all, increased computerization will bring more, rather than less, complex demands on the NCO.

2. How effectively would you expect this soldier to meet these future NCO requirements?

Not likely to meet the NCO demands described under these conditions.		Likely to be generally successful, but will struggle to meet the NCO demands described under these conditions.			Likely to successfully meet or exceed NCO demands described under these conditions.	
LOW		**MODERATE**			**HIGH**	
1	**2**	**3**	**4**	**5**	**6**	**7**

Scenario #3: Increased Scope of Technical Skill Requirements

The future Army will be based on a combination of advanced weapons systems, various levels of information systems, and sophisticated communications. Organizationally, a significant part of the Army is intended to contain small, flexible battle force teams. These teams will be highly trained with a mixing of roles across ranks and with all team members cross-trained in each others' skills. The existing structure of a large number of specialized MOS likely will be replaced by a system in which NCOs are classified into broad areas of job abilities based primarily on types of units or echelons of employment. NCOs in battle forces will be expected to employ a full array of organic and supporting fires, maneuver and transportation, intelligence gathering facilities, engineering methods, data communications, and protective measures. Logistics, including supply, maintenance and repair, and field medical and evacuation will become organic requirements of the battle force. The NCO of the future will have almost unlimited access to information sources for diagnoses and step-by-step procedures, but actual performance will still have to be learned and practiced. The end result will be an increase in the technical requirements for future NCOs, probably doubling or tripling the number of skill tasks associated with today's NCOs.

3. How effectively would you expect this soldier to meet these future NCO requirements?

Not likely to meet the NCO demands described under these conditions.		Likely to be generally successful, but will struggle to meet the NCO demands described under these conditions.			Likely to successfully meet or exceed NCO demands described under these conditions.	
LOW		**MODERATE**			**HIGH**	
1	**2**	**3**	**4**	**5**	**6**	**7**

Scenario #4: Increased Requirements for Broader Leadership Skills at Lower Levels

Over the next 20 years, broader leadership skills will be a critical requirement of the NCO. Units the size of current platoons and companies will be the focal points of operations. Combat support and combat service support organizations will be even smaller with only 1 to 5 person cells providing specialized assistance. It will be common for units to be widely scattered and, while communication and information linkage will increase, there will be less physical contact between units of all sizes. In many situations the chain of command will be temporary and will be through information linkages rather than established relationships. Furthermore, because many missions will be situation specific, NCOs will not be able to rely as much on past experiences when making decisions in new situations.

As a result, many of the requirements for leadership, decision making, initiative, responsibility, and accountability that are today thought of as company-grade and junior officer requirements will become the domains of the E7 and E6. In turn, the level of leadership, authority, and responsibility that is currently associated with platoon sergeants, staff shift supervisors, detachment, and shop supervisors will migrate down to the E5 and E4 levels. Although at some point, future NCOs will be able to access automated decision matrices or artificial intelligence to assist them with their leadership decisions, they will have many requirements similar to what leaders have always faced – unpredicted situations, human interactions and stresses, system malfunctions, and time pressures. The difference will be that these requirements, and their consequences, will be experienced in a greater degree and at lower ranks by future NCOs.

4. How effectively would you expect this soldier to meet these future NCO requirements?

Not likely to meet the NCO demands described under these conditions.		Likely to be generally successful, but will struggle to meet the NCO demands described under these conditions.			Likely to successfully meet or exceed NCO demands described under these conditions.	
LOW		**MODERATE**			**HIGH**	
1	**2**	**3**	**4**	**5**	**6**	**7**

Scenario #5: Need to Manage Multiple Operational Functions and Deal with the Inter-relatedness of Units

The future Army will have a less rigid organizational structure, more mission type operations that have multiple purposes (e.g., mixed peace making/peacekeeping), more independent operations at lower levels, and increased low-level lethality. It will still employ the engagement systems of maneuver; fire support; information dominance; reconnaissance, surveillance, and intelligence; mobility and survivability; and air defense along with the integrating systems of command and control and combat service support. However, as technology and information flow improves, these will be planned for, integrated, and executed at lower and lower levels. With more capabilities at lower levels and operating under mission-type orders, NCOs will have more flexibility in the courses of actions available to them in any given situation. Along with this will come a requirement to be more aware of how one's own actions affect the total environment in which the NCO is operating. Impacts on other units, higher headquarters missions, civilian populations, strategic goals, and fratricide possibilities must be weighed by individual NCOs into any course of action they are contemplating. The ability to predict the effects of an activity onto others within the battlespace will become a crucial element of NCO-led operations. The boundaries of these operations will not be limited to what they can see or even by physical limits. NCOs must be able to operate by projecting the effects of their decisions in many directions and levels simultaneously. Although these requirements will be accompanied by improvements in technology and decision software, the timing and control of the use of available systems will remain very much a human element.

5. How effectively would you expect this soldier to meet these future NCO requirements?

Not likely to meet the NCO demands described under these conditions.		Likely to be generally successful, but will struggle to meet the NCO demands described under these conditions.			Likely to successfully meet or exceed NCO demands described under these conditions.	
LOW		**MODERATE**			**HIGH**	
1	**2**	**3**	**4**	**5**	**6**	**7**

Scenario #6: Mental and Physical Adaptability and Stamina

There is no indication that the current demands for physical strength and endurance will change much in the near future. However, future operations will likely involve new aspects of physical, psychomotor, and mental skills. Future conflicts are expected to involve more intense and sustained operations that will require enough physical and mental stamina to conduct high paced operations over long periods. Individuals must become capable of cycling between periods of work and rest instantaneously and at unpredictable intervals. Mental sharpness will be important as individuals will be required to process, sort, and prioritize digital information and data flow without being overwhelmed, even when fatigued or stressed. NCOs must be able to recognize and respond to mental cues and images (such as icons and graphics) rather than visual or sound stimuli of real-life events.

In these intense fluid situations, NCOs must be capable of solving problems effectively without knowing all of the facts. Operations in uncertain environments will demand that NCOs are able to make reasoned, logical assessments of conditions without exaggerating the situation or becoming distressed. Situations will change rapidly and NCOs will often acquire information *en route*. Equipment failures, fluidity of operations, and novel missions will demand frequent and sometimes unprecedented levels of mental and physical adaptability to changing conditions.

6. How effectively would you expect this soldier to meet these future NCO requirements?

Not likely to meet the NCO demands described under these conditions.		Likely to be generally successful, but will struggle to meet the NCO demands described under these conditions.			Likely to successfully meet or exceed NCO demands described under these conditions.	
LOW		**MODERATE**			**HIGH**	
1	**2**	**3**	**4**	**5**	**6**	**7**

Please use the answer sheet to rate how confident you are about the accuracy of the ratings you have provided.

Appendix C
Experience and Activities Record

Experience & Activities Record

MARKING INSTRUCTIONS

- Use a No. 2 pencil only.
- Do not use ink, ballpoint, or felt tip pens.
- Make solid marks that fill the response completely.
- Erase cleanly any marks you wish to change.
- Make no stray marks on this form.

CORRECT: ● **INCORRECT:** ✓ ✗ ◌ ▬

ID Number

0	0	0	0
1	1	1	1
2	2	2	2
3	3	3	3
4	4	4	4
5	5	5	5
6	6	6	6
7	7	7	7
8	8	8	8
9	9	9	9

This form lists a variety of experiences, activities, or assignments some soldiers have had. Please respond to each item based on your experience.

Frequency
In the last 2 years, how often have you performed each activity?

◆ Experiences and Activities	Never	A few times a year	About once a month	A few times a month	A few times a week	Daily
Computer Related Activities						
1. Used a PC, Mac, or laptop.	1	2	3	4	5	6
2. Communicated using e-mail.	1	2	3	4	5	6
3. Used the Internet for job or training requirements.	1	2	3	4	5	6
4. Used the Windows NT operating system.	1	2	3	4	5	6
5. Operated an Army-specific computer system (e.g., IVIS, ASAS, FBCB2, AFATDS).	1	2	3	4	5	6
6. Troubleshooted a computer system malfunction.	1	2	3	4	5	6
7. Used Windows Office programs to do job tasks (e.g., Word®, Access®, Excel®, PowerPoint®).	1	2	3	4	5	6
8. Trained or assigned as an instructor/operator (I/O) on any computer based simulator (e.g., COFT, BBS, CBS, SIMNET, Janus).	1	2	3	4	5	6
Leadership/Supervisory						
9. Assigned to duty position with a responsibility for supervising 2 or more soldiers.	1	2	3	4	5	6
10. Provided performance feedback to subordinates.	1	2	3	4	5	6
11. Established goals or other incentives to motivate subordinates.	1	2	3	4	5	6
12. Corrected unacceptable conduct of a subordinate.	1	2	3	4	5	6
13. Trained other soldiers in a task or procedure.	1	2	3	4	5	6
14. Conducted formal inspection of subordinates' completed work.	1	2	3	4	5	6
15. Counseled subordinates regarding career planning.	1	2	3	4	5	6
16. Counseled subordinates with disciplinary problems.	1	2	3	4	5	6
17. Served as a member of a unit advisory council or committee.	1	2	3	4	5	6
18. Applied and supervised all 8 steps of troop leading procedures (TLP).	1	2	3	4	5	6
Additional Duties						
19. Volunteered for additional duties/assignments.	1	2	3	4	5	6
20. Requested additional training opportunities.	1	2	3	4	5	6

◆ Assignments and Positions

	Duration: How much time have you spent in each of the following?				
	Never	Less than 6 months	6 months to a year	1 year to 2 years	More than 2 years
Duration of Experiences					
21. Total time spent in duty position one grade higher than actual grade.	1	2	3	4	5
22. Total time spent in a leadership or supervisory position.	1	2	3	4	5
23. Total time spent in MTOE slot assignment.	1	2	3	4	5
24. Total time in a unit specialty assignment (e.g., Commander's or First Sergeant's driver, Assistant Training NCO, NBC, Unit Lifesaver).	1	2	3	4	5

◆ Training and Duties

	Frequency: How many times have you done each of the following?		
	Never	Once	Twice or more
Formal Training/Assignments			
25. Participated in CTC/NTC/JRTC rotation or FTX over 30 days.	1	2	3
26. Deployed on combat mission.	1	2	3
27. Deployed on peace-keeping mission.	1	2	3
28. Prepared a lesson plan.	1	2	3
29. Led a PT class.	1	2	3
30. Taught a platform class to 5 or more people.	1	2	3
31. Served as an assistant instructor in a class of 10 or more people.	1	2	3
32. Been part of a crew to perform Table VIII, Table XII, or TCPC.	1	2	3
33. Participated as a team leader or above in a live fire exercise (LFX).	1	2	3
34. Conducted primary marksmanship instruction (PMI).	1	2	3
Communications			
35. Received and implemented a written operations order.	1	2	3
36. Issued a 5 paragraph oral operations order.	1	2	3
37. Prepared and submitted a written report of recognition for a subordinate.	1	2	3
38. Prepared and conducted a briefing for 2 or more officer, senior NCO, or civilian personnel.	1	2	3
39. Prepared a written plan/schedule of future subordinate activities covering 5 days or more.	1	2	3
40. Prepared a written counseling statement.	1	2	3
Inspections, Drills and Ceremonies, Official Duties			
41. Led/commanded soldiers in drill and ceremony activities.	1	2	3
42. Conducted an inspection in ranks or standby.	1	2	3
43. Performed as Color Guard.	1	2	3
44. Acted as assistant commander at funeral detail or other public ceremony.	1	2	3
45. Served as a VIP escort.	1	2	3
46. Appeared before a Soldier of the Month (or equivalent) Board.	1	2	3

Appendix D

Personnel File Form-21

Personnel File Form-21

MARKING INSTRUCTIONS

- Use a No. 2 pencil only.
- Do not use ink, ballpoint, or felt tip pens.
- Make solid marks that fill the response completely.
- Erase cleanly any marks you wish to change.
- Make no stray marks on this form.

CORRECT: ● INCORRECT: ✓ ✗

ID Number			
0	0	0	0
1	1	1	1
2	2	2	2
3	3	3	3
4	4	4	4
5	5	5	5
6	6	6	6
7	7	7	7
8	8	8	8
9	9	9	9

◆ Awards/Commendations

1. Mark the awards and decorations listed below that you have received. If you have received any awards or decorations not listed, mark "other" and specify the name of the award or decoration.

- Soldier's Medal or higher award
- Bronze Star Medal (Valor or Merit)
- Defense Meritorious Service Medal
- Meritorious Service Medal
- Air Medal (Valor or Merit)
- Joint Service Commendation Medal
- Joint Achievement Medal
- Purple Heart
- Combat Infantry Badge
- Combat Field Medical Badge
- Expert Infantry Badge
- Expert Field Medical Badge
- Basic Parachutist Badge
- Senior Parachutist Badge
- Master Parachutist Badge
- Divers Badge
- Explosive Ordnance Disposal Badge
- Pathfinder Badge
- Aircraft Crewman Badge
- Nuclear Reactor Operator Badge
- Ranger Tab
- Special Forces Tab
- Driver and Mechanic Badge
- Air Assault Badge
- Drill Sergeant Identification Badge
- US Army Recruiter Badge
- Campaign Star (Battle Star)
- Equivalent awards and decorations earned in other US uniformed services
- Army Reserve Components Acheivement Medal
- Southwest Asia Medal
- Other _____
- Other _____

If you received any of the following medals, indicate how many.	1	2	3 or more
Army Commendation Medal (Valor or Merit)	○	○	○
Army Achievement Medal	○	○	○
Good Conduct Medal	○	○	○

Military Academic Achievement

- Distinguished Honor Graduate
- Distinguished Leadership Award
- Commandant's List

Military Board Achievement

- Soldier/NCO of the Quarter - Brigade Level
- Soldier/NCO of the Year - Brigade Level
- Soldier/NCO of the Quarter - Installation/Division Level
- Soldier/NCO of the Year - Installation/Division Level
- Soldier/NCO of the Year - MACOM Level

2. How many <u>Memoranda/ Letters</u> of Appreciation, Commendation, Achievement have you received...

Write the number in the boxes.

Then, fill in the matching circle below each box.

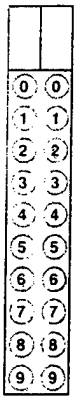
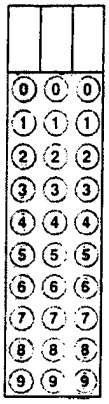
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9

3. How many <u>Certificates</u> of Achievement have you received...

0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9

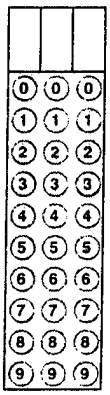
◆ Military Education

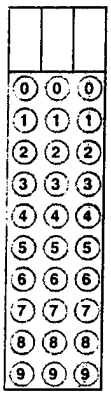
4. Indicate courses listed below that you have successfully completed. Do not include BT, OSUT, or AIT.

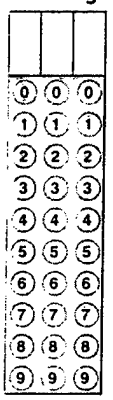
- ○ PLDC
- ○ Airborne School
- ○ BNCOC - If yes, how many weeks? → 
- ○ NBC School
- ○ Ranger School
- ○ Air Assault School
- ○ Special Forces Qualification Course
- ○ Any other course of at least 40 hours duration - If yes, how many? → 
- ○ Military correspondence course credit hours - If yes, how many? → 
- ○ EMT Basic Certification
- ○ EMT Intermediate Certification
- ○ EMT Paramedic Certification

◆ Civilian Education

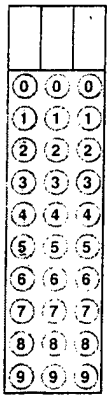
5. List the total number of semester hours you have earned since you have been on active duty.

a. Career/ Trade School 

b. Vo Tech 

c. College 

Of the semester hours you have earned since you have been on active duty, indicate how many were paid for through the Army's Tuition Assistance Program.

a. Career/ Trade School 

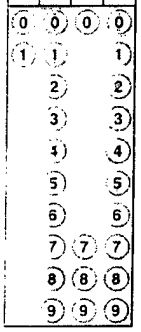
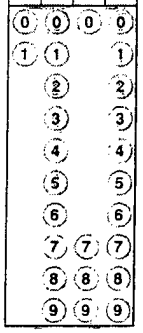
b. Vo Tech 

c. College 

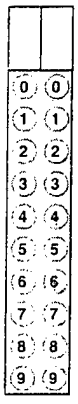
6. Have you earned a civilian college degree since you have been on active duty?

- ○ Yes - If yes, indicate the type of degree(s)
  - ○ Associates
  - ○ Bachelors
  - ○ Masters
  - ○ Other _____
- ○ No

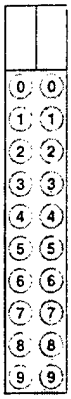
If you answered yes to Question 6, indicate when you started to work on your degree and when you completed it.

Started		Finished	
Mo.	Yr.	Mo.	Yr.
			

◆ Disciplinary Action

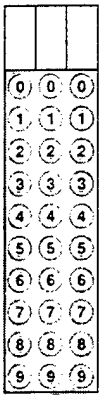
7. How many Articles 15 have you received... 

8. How many Flag Actions (i.e., suspension of favorable personnel action) have you received...



◆ Test Scores

9. What was your last Physical Readiness Test score? (score ranges from 0-300)



10. What was your last Weapon Qualification?

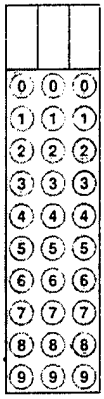
- ○ Unqualified
- ○ Marksman (MKM)
- ○ Sharpshooter (SPS)
- ○ Expert (EXP)

11. Have you retaken the ASVAB since your initial enlistment screening?

- ○ Yes - If yes, how many times have you retaken the ASVAB/AFCT exam? →

- ○ No

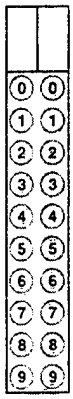
12. What is your current General Technical (GT) score of record?



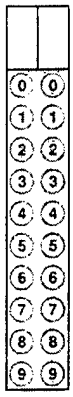
◆ ACES Participation

This section asks about your participation in programs sponsored by the Army Continuing Education System (ACES).

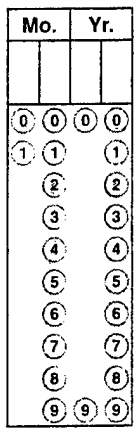
13. How many MOS Improvement/ Soldier (Unit) Training Courses sponsored by Army Education have you successfully completed?



14. a. How many Army Education NCO Leadership Development Courses did you successfully complete prior to being promoted to your current grade?



b. When did you complete the last NCO Leadership Development Course prior to being promoted to your current grade?



- ○ Not applicable

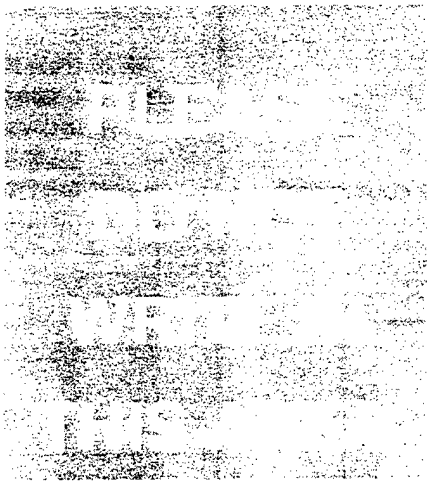
Please continue on the next page.

15. **To what extent have Army Education programs such as Tuition Assistance, college/vocational-technical courses, NCO Leadership Development Courses, and MOS Improvement Courses improved your competence to perform at the next higher grade level?**

- ○ Does not apply; I have not participated in any Army Education programs.
- ○ Army Education programs have not improved my competence.
- ○ Army Education programs have slightly improved my competence.
- ○ Army Education programs have somewhat improved my competence.
- ○ Army Education programs have greatly improved my competence.

16. **To what extent have Army Education programs enhanced your performance as a soldier?**

- ○ Does not apply; I have not participated in any Army Education programs.
- ○ Army Education programs have not enhanced my performance.
- ○ Army Education programs have slightly enhanced my performance.
- ○ Army Education programs have somewhat enhanced my performance.
- ○ Army Education programs have greatly enhanced my performance.



Appendix E
Semi-Structured Interview Rating Scales

Adaptability

Can modify behavior or plans as necessary to reach goals or to adapt to changing goals. Is able to maintain effectiveness when environments, tasks, responsibilities, or personnel change. Easily commits to learning new things when the technology, mission, or situation requires it.

1 2	3 4 5	6 7
LOW	**MODERATE**	**HIGH**
• Floundered when it was necessary to think quickly on his or her feet.	• Developed workable solutions to problems, changing direction when it was necessary.	• Developed effective solutions to problems, switching from one situation to another quickly and smoothly.
• Could not learn new procedures when dealing with crisis or unfamiliar situations.	• Dealt with crisis or unfamiliar situations in a reasonably effective manner by following learned procedures.	• Learned new procedures quickly and effectively as required by crisis or unfamiliar situations.
• Resisted learning new ways of doing things when the technology was available or the situation required it.	• Modified behavior and learned new ways of doing things as the technology or situation required, but did so with hesitation or difficulty.	• Demonstrated exceptional ability to implement new procedures as technology or situations warranted.

Self-Management and Self-Directed Learning Skill

Manages the full range of own career, personal, and financial responsibilities through strategies such as setting long- and short-term goals. Identifies personal training needs, plans education and training experiences to meet them, and evaluates own training success. Seeks to continually develop job and personal skills by participating in self-study, reading, training programs, and/or educational classes. Uses effective personal learning strategies. Works effectively without direct supervision, but seeks help and advice from others when appropriate.

1 2	3 4 5	6 7
LOW	**MODERATE**	**HIGH**
• Demonstrated poor work performance because of ineffective management of personal responsibilities.	• Effectively balanced work assignments and personal responsibilities.	• Established priorities and plans to handle work and personal life. Was able to use these to avoid conflicts.
• Suffered financial trouble due to poor planning and irresponsible behavior.	• Was able to meet most short-term financial responsibilities, but has not set long-term financial goals.	• Was able to meet financial responsibilities by setting both short- and long-term goals.
• Did not take initiative to pursue training courses or other learning opportunities.	• Pursued additional course/training when supervisor advised him/her to do so.	• Actively pursued additional training to improve job skills and increase chance of promotion.
• Could not evaluate personal training needs.	• Was usually able to evaluate personal training needs.	• Identified personal training needs to meet future requirements (i.e., didn't wait for problem to arise; was proactive in training).

Level of Effort and Initiative on the Job

Demonstrates high effort in completing work. Takes independent action when necessary. Seeks out and willingly accepts responsibility and additional challenging assignments. Persists in carrying out difficult assignments and responsibilities.

1 2	3 4 5	6 7
LOW	**MODERATE**	**HIGH**
• Showed little willingness to take on challenging assignments. • Failed to put in the extra time or effort necessary to complete a job. • Did not attempt to try a difficult task/assignment.	• Displayed willingness to take responsibility for completing challenging assignments. • Put in additional time and effort when it was necessary to complete a job. • Tried to complete a difficult assignment, but eventually gave up.	• Showed initiative and independence in taking on challenging assignments. • Willingly put in extraordinary time and effort to ensure effective completion of work. • Persevered in completing difficult assignments.

Level of Integrity and Discipline on the Job

Maintains high ethical standards. Does not succumb to peer pressure to commit prohibited, harmful, or questionable acts. Demonstrates trustworthiness and exercises effective self-control. Understands and accepts the basic values of the Army and acts accordingly.

1 2	3 4 5	6 7
LOW	**MODERATE**	**HIGH**
• Blamed others for his or her own job-related errors.	• Took some responsibility for a job-related mistakes, but made some excuses about why the mistake was made to minimize the error.	• Admitted to a job-related error and made sure others were not blamed for it.
• Looked the other way, even when laws or regulations were being violated.	• Confronted or reported others when they were committing a serious wrong, but looked away in less serious situations or when there was less chance of being caught.	• Held self and others to the highest standard of ethics, regardless of consequences.
• Demonstrated self-indulgent behavior.	• Controlled self-indulgent impulse, but gave in when there was little of no chance of being caught.	• Did not succumb to peer pressure (and was not accepting of others who did so), even though there was little or no chance of being caught.

Relating to and Supporting Peers

Treats peers in a courteous, respectful, and tactful manner. Provides help and assistance to others. Backs up and fills in for others when needed. Works effectively as a team member.

1 2	3 4 5	6 7
LOW	**MODERATE**	**HIGH**
• Was disrespectful to peers and disregarded their opinions or concerns. • Hung on adamantly and aggressively to own position, refusing to compromise. • Refused to help peers or fill in for others when needed.	• Tried to be respectful and courteous when dealing with peers. • Accepted compromise when it was offered or proposed compromises when there was an obvious disagreement. • Agreed to assist peers or fill in, but only did so when asked.	• Was clearly courteous, respectful, and tactful in dealing with peers. • Offered a compromise to prevent a disagreement among peers that helped keep a task on track. • Willingly assisted or filled in for peers when needed.

Leadership Skills/Potential

Inspires, motivates and guides others toward goal accomplishment; adapts leadership style to fit a variety of situations. Recognizes, encourages, and rewards effective performance of individual subordinates or a team. Communicates team goals and shares relevant information with team members. Evaluates and identifies individual or unit training needs and institutes formal/informal programs to address them.

1 2	3 4 5	6 7
LOW	**MODERATE**	**HIGH**
• Unsuccessful at motivating subordinates.	• Generally able to motivate subordinates to accomplish difficult or unpleasant job assignments.	• Consistently motivates peers and subordinates to accomplish difficult or unpleasant job assignments.
• Provided negative feedback to a subordinate that did not result in improved performance (e.g., discipline was too harsh).	• Provided constructive developmental feedback that led to some performance improvement.	• Provided constructive feedback to correct a problem and motivate a subordinate while helping the individual maintain a high level of self-esteem.
• Failed to develop a training plan or provide activities that would enhance learning (e.g., demonstrate tasks, provide practice and feedback).	• Developed a simple training plan to guide training presentation and included some activities to reinforce learning (e.g., demonstration, practice, feedback).	• Developed effective training plans to structure material and used a wide variety of techniques to ensure learning (e.g., demonstration, practice, feedback, testing, and remedial training).
• Did not recognize or reward effective teamwork.	• Recognized effective teamwork, but was slow to do so or only after it was pointed out.	• Recognized and rewarded effective teamwork in a way that fostered team unity.

MOS/Occupation-Specific Knowledge and Skill

Possesses the necessary knowledge and skill to perform MOS/occupation-specific technical tasks at the appropriate skill level. Stays informed of the latest developments in field.

1 2	3 4 5	6 7
LOW	**MODERATE**	**HIGH**
• Did not display knowledge required to perform routine assignments or tasks in MOS at current skill level. • Unaware of recent developments relevant to his or her MOS. • Incorrectly responded to questions about his or her MOS/occupation.	• Displayed good knowledge of most aspects of the MOS/occupation at current skill level. • Knowledgeable of some new developments in his or her MOS/occupation, but not the most recent or critical ones. • Correctly responded to most questions relevant to his or her MOS/occupation, but appeared to be guessing; soldier incorrectly responded to a question about his or her job, but the answer was not too far off-the-mark.	• Displayed in-depth knowledge of MOS/occupation above current skill level. • Knowledgeable of recent developments relevant to his or her MOS/occupation. • Correctly and confidently responded to questions relevant to the knowledge required in his or her MOS/occupation.

Oral Communication Skill

Speaks in a clear, organized, and logical manner. Communicates detailed information, instructions, or questions in an efficient and understandable way. Note that this skill refers to how well the individual can speak and communicate, not whether technical expertise is high or low.

1 2	3 4 5	6 7
LOW	**MODERATE**	**HIGH**
• Inappropriately used non-verbal cues such as eye contact, facial expressions, and hand gestures. • Responded to questions in a disjointed and unconfident manner; strayed to irrelevant topics. • Used poor grammar and spoke in an awkward or confusing manner.	• Helped communicate points by using non-verbal cues such as eye contact, facial expressions, and hand gestures. • Expressed self clearly and focused on relevant issues, but lacked confidence. • Spoke clearly with only a few minor stammers or grammatical errors.	• Effectively communicated points using non-verbal cues such as eye contact, facial expressions, and hand gestures. • Expressed self, concisely, clearly, confidently, and persuasively, and focused on pertinent issues. • Spoke fluently and articulately, using appropriate grammar.

Military Presence

Presents a positive and professional image of self and the Army even when off duty. Maintains proper military appearance.

1 2	3 4 5	6 7
LOW	**MODERATE**	**HIGH**
• Soldier was dressed sloppily or improperly. • Displayed bad posture or poor military bearing throughout the interview. • Appeared to be unfit or overweight (e.g., above the Army weight standard for age and height).	• Dressed relatively neatly and properly, maintaining basic Army standards. • Displayed good military posture and bearing during most of the interview. • Appeared to be fit and within Army weight standards.	• Dressed neatly and sharply, exceeding basic Army standards. • Maintained excellent military posture and bearing throughout the interview. • Appeared to be exceptionally fit and athletic and was well within Army weight standards.