40 Project A

12 Years of R & D

John P. Campbell and Deirdre J. Knapp

This chapter[1] is about personnel selection and classification research on a scale never before attempted in terms of (a) the types and variety of information collected, (b) the number of jobs that were considered simultaneously, (c) the size of the samples, and (d) the length of time that individuals were followed as they progressed through the organization.

The effort, commonly known as Project A, was sponsored by the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI). For contract management reasons the research program was conducted as two sequential projects: Project A (1982–1989) and Career Force (1990–1994), which worked from a single overall design (described subsequently).

Collectively, these projects attempted to evaluate the selection validity and classification efficiency of systematically sampled domains of prediction information for different selection and classification goals for the entire enlisted personnel system of the U.S. Army, using various alternative decision rules (i.e., "models"). Pursuing such ambitious objectives required the development of a comprehensive battery of new tests and inventories, the development of a wide variety of training and job performance measures for each job in the sample, four major worldwide data collections involving thousands of Army enlisted job incumbents for one to two days each, and the design and maintenance of the resulting database.

The truly difficult part was the never-ending need to develop a consensus among all of the project participants regarding literally hundreds of choices among measurement procedures, analysis methods, and data collection design strategies. Although many such decisions were made in the original design stage, many more occurred continuously as the projects moved forward, driven by the target dates for the major data collections, which absolutely could not be missed. The fact that all major parts of the projects were completed within the prescribed time frames and according to the specified research design was a source of wonder for all who participated.

ORIGINS OF PROJECT A

Project A resulted from pressing organizational needs from within the Department of Defense (DoD), and not from unsolicited proposals for research grants. The post-Vietnam development of the all volunteer force, the comprehensive force modernization programs of the services during the 1970s, periods of high attrition, and the fallout from the misnorming of forms 6/7 of the Armed Services Vocational Aptitude Battery (ASVAB) all created great pressure on the military services to make certain that their selection and classification procedures were as valid as possible. The culmination was a 1980 Congressional mandate to conduct a more thorough validation of ASVAB for selection and classification purposes. Prior to 1980, virtually all validation evidence was based

[1] Much of this chapter was drawn from Campbell and Knapp (2001) and from the technical reports generated by Project A and Career Force. We are indebted to all of the individuals from the Army Research Institute (ARI), the Human Resources Research Organization (HumRRO), the American Institutes for Research (AIR), the Personnel Decisions Research Institute (PDRI), and the University of Minnesota who contributed their technical expertise, initiative and effort, personal discipline, and peer leadership to these projects. It was the experience of a lifetime.

on training grades as criteria. The DoD's formal response was the Joint-Service Job Performance Measurement/Enlistment Standards Project (JPM).

Project A was the Army's contribution to JPM. Moreover, the Army viewed the Congressional mandate as an opportunity to address a much larger set of personnel research questions than just the validation of ASVAB against job performance. In September 1982, a contract for Project A was signed with the Human Resources Research Organization (HumRRO) and its subcontractors, the American Institutes for Research (AIR) and Personnel Decisions Research Institute, Inc. (PDRI).

The overall design of the Project A program was intended to be fundamentally different from the conventional paradigm that dominated personnel research from 1906 to 1982, which consisted of computing the correlation between a single predictor score, or a predictor composite score, and a single criterion measure of performance obtained from a sample of job incumbents. Literally thousands of such estimates have been generated over the past 100+ years (e.g., Ghiselli, 1973; Hunter & Hunter, 1984; Nathan & Alexander, 1988; Schmidt, 1988; Schmidt & Hunter, 1998; Schmitt, Gooding, Noe, & Kirsch, 1984).

There are probably legitimate reasons why single investigators working to generate one bivariate distribution at a time has served as the dominant paradigm through most of our history. For one, the recurring problem of how best to select individuals for a particular job in a particular organization is a very real one, and a rational management will devote resources to solving such a problem. Also, certain structural and technological factors have worked against the establishment of long-term coordinated research projects that dealt with large parts of the personnel system at one time. For example, the field of industrial and organizational psychology is not very large and the supply of research labor is limited. When the basic outline of Project A/Career Force was proposed, there was no single organization or university group that had the resources necessary to carry it out. Coalitions of organizations had to form. Also, until fairly recently, there were no means available for coordinating the efforts of researchers who are geographically scattered. Neither was there a technology for acquiring and maintaining a large central database that could be accessed and analyzed efficiently from remote locations.

In general, the dominant paradigm came to be so because of the constraints imposed by technology, the structural characteristics of the research enterprise itself, and the contingencies built into the reward structures for individual investigators.

ENABLING OF PROJECT A

Fortunately, along with the Army's need to address enlisted selection and classification as a system, there were concomitant developments in the structure and technology of the personnel research enterprise. For example, advances in computerized database management and electronic communication made it possible to design, create, edit, update, and maintain a very large database that could be accessed from anywhere. What is routine now was new and liberating in 1982.

Advances in computerization also permitted use of new testing technologies, and the development of powerful, linear programming algorithms made the estimation of classification efficiency and the comparison of alternative selection/classification strategies using the entire Army database a very manageable analytic problem. Certainly, the development of confirmatory techniques within the general domain of multivariate analysis models opened up several powerful strategies for generalizing research findings from a sample of jobs to a population of jobs and from the specific measures that were used to a latent structure of constructs.

Finally, the realization in industrial and organizational psychology during the 1970s that one of our fundamental tasks is to learn things about an appropriately defined population, and not to learn more and more specific things about specific samples, changed the field's approach to the estimation of selection validity and classification efficiency. Meta-analysis and corrections for attenuation and restriction of range were no longer novel games to play. They were a necessary part of statistical estimation.

In sum, the intent was to design a research program that would be directly useful for meeting the system's needs, both as they existed initially and as changes took place. Simultaneously, everyone hoped that by considering an entire system and population of jobs at once, and by developing measures from a theoretical/taxonomic base, the science of industrial and organizational psychology would also be served.

SPECIFIC RESEARCH OBJECTIVES

The objectives were ambitious and spanned a continuum from operational/applied concerns to more theoretical interests. They are summarized as follows:

1. Identify the constructs that constitute the universe of information available for selection/classification into entry-level skilled jobs, given no prior job experience, and develop predictor measures for those constructs identified as "best bets."
2. Develop multiple measures of entry-level and noncommissioned officer (NCO) job performance.
3. Develop a general model of performance for entry-level skilled jobs and for NCO jobs.
4. Validate existing selection measures (i.e., ASVAB) against training and job performance.
5. On the basis of the "best bet" constructs, validate a battery of new predictor measures.
6. Estimate validity and incremental validity of the new predictors.
7. Estimate the degree of differential prediction across (a) major domains of predictor information, (b) major factors of job performance, and (c) different types of jobs.
8. Develop new analytic methods to evaluate optimal selection and classification.
9. Compare the marginal utility of performance across jobs.
10. Develop a fully functional research database that includes all archival research data on the three cohorts of new Army accessions included in the research program.

OVERALL RESEARCH DESIGN

The first 6 months of the project were spent developing a final detailed version of the operative research plan, which was published in 1983 as ARI Research Report No. 1332, *Improving the Selection, Classification, and Utilization of Army Enlisted Personnel: Project A Research Plan.*

SAMPLING JOBS (MOS)

In 1982 the population of enlisted jobs included approximately 275 different Military Occupational Specialties (MOS), and the entire enlisted force was approximately 800,000. Because data could not be collected from all of them, MOS were sampled representatively after considering the tradeoff between the number of jobs to be researched and the number of individuals sampled from each job. Cost considerations dictated that 18–20 MOS could be studied if the initial goal was 500 job incumbents per job, and this assumed that a full array of job-specific performance measures would be developed for only a subset of those MOS.

An initial sample of 19 MOS was drawn on the basis of the following considerations:

1. High-density jobs that would provide sufficient sample sizes
2. Representation of the Career Management Fields (CMF) into which MOS are organized
3. Representation of the jobs judged most crucial to the Army's missions

The initial set of 19 MOS included only 5% of Army jobs but represented 44% of the soldiers recruited in FY81. An independent (i.e., without considering the CMF designation) cluster analysis of MOS (based on task content similarity) was carried out via Army subject matter expert (SME)

TABLE 40.1
Sampled Military Occupational Specialties (MOS)

	MOS Batch A		MOS Batch Z
11B	Infantryman	12B	Combat engineer
13B	Cannon crewman	16S	MANPADS crewman
19E/K	Armor tank crewman	27E	Tow/dragon repairer
31C	Single-channel radio operator	29E	Communications-electronics radio repairer
63B	Light-wheel vehicle mechanic	51B	Carpentry/masonry specialist
71L	Administrative specialist	54B	NBC specialist
88M	Motor transport operator	55B	Ammunition specialist
91A/B	Medical specialist/medical NCO	67N	Utility helicopter repairer
95B	Military police	76Y	Unit supply specialist
		94B	Food service specialist
		96B	Intelligence analyst

judgments to evaluate the representativeness of the sample of 19 and to make changes in the composition of the sample if it was judged appropriate to do so. Two jobs were added, and Table 40.1 shows the resulting MOS ($n = 21$) that were studied over the course of Project A. “Batch A” MOS received the most attention in that soldiers in these jobs were administered a full array of first- and second-tour job performance measures, including hands-on work sample tests, written job knowledge tests, and Army-wide and MOS-specific ratings. Soldiers in “Batch Z” were not measured as extensively with regard to the job performance criterion measures.

Data Collection Design

The basic design framework and major samples are depicted in Figure 40.1. The design encompassed two major cohorts, each of which was followed into their second tour of duty (i.e., enlistment term) and collectively produced six major research samples. Development of the predictor and criterion measures administered during the major phases of this research involved dozens of additional smaller data collection efforts as well, for purposes of pilot and field-testing. Each of the six major data collections is briefly characterized below.

Concurrent Validation (CVI) Sample

This sample was drawn from soldiers who had entered the Army between July 1983 and June 1984 and had served 18–24 months. Data were collected from soldiers and their supervisors at 13 posts in the United States and at multiple locations in Germany. Batch A soldiers (see Table 40.1) were assessed for 1.5 days on the first-tour job performance measures and for a half-day on the new predictor measures. Batch Z soldiers were tested for a half-day on a subset of the performance measures and a half-day on the new predictors.

Longitudinal Validation Predictor (LVP) Sample

Virtually all new recruits who entered the Army into one of the sampled MOS from August 1986 through November 1987 were assessed on the 4-hour Experimental Battery (to be described) within 2 days of first arriving.

Longitudinal Validation End-of-Training (LVT) Sample

End-of-training performance measures were administered to those individuals in the LVP sample who completed advanced individual training (AIT), which could take from 2–6 months, depending on the MOS.

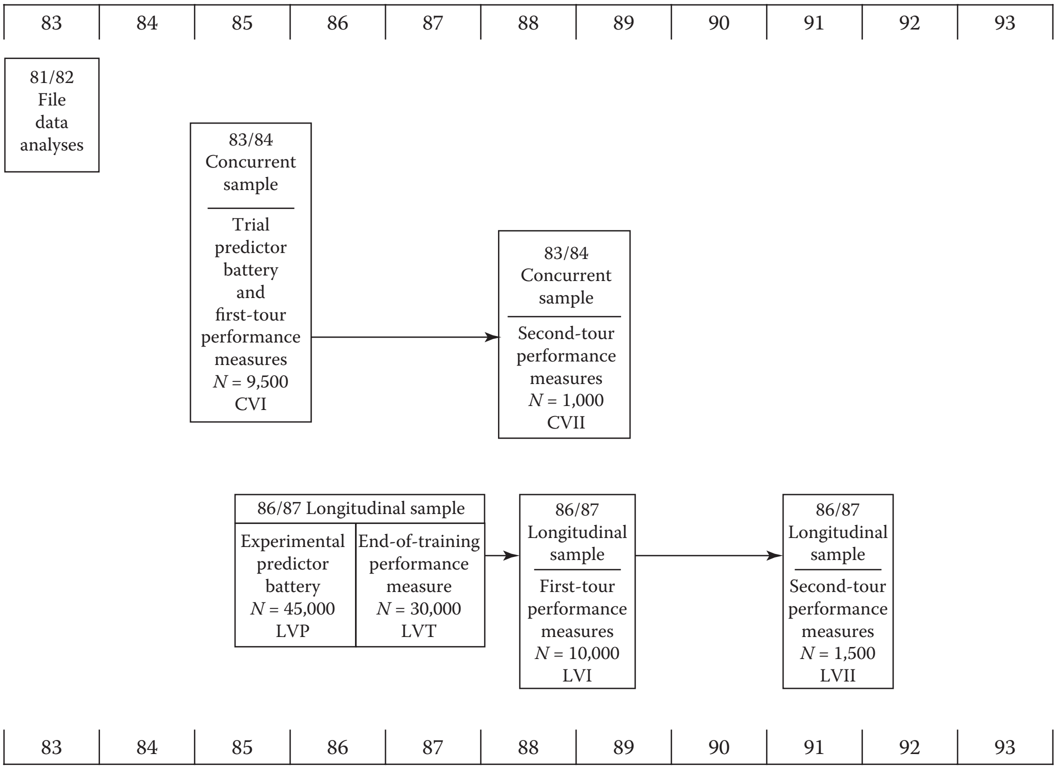
83
84
85
86
87
88
89
90
91
92
93
81/82 File data analyses
83/84 Concurrent sample
Trial predictor battery and first-tour performance measures N = 9,500 CVI
83/84 Concurrent sample
Second-tour performance measures N = 1,000 CVII
86/87 Longitudinal sample
Experimental predictor battery N = 45,000 LVP
End-of-training performance measure N = 30,000 LVT
86/87 Longitudinal sample
First-tour performance measures N = 10,000 LVI
86/87 Longitudinal sample
Second-tour performance measures N = 1,500 LVII

FIGURE 40.1 Project A: Overall data collection design and major research samples.

Longitudinal Validation First-Tour (LVI) Sample

Individuals in the 86/87 cohort who were measured with the Experimental Battery, completed training, and remained in the Army were assessed with the first-tour job performance measures after serving 18–24 months. Data were collected from 13 posts in the United States and multiple locations in Europe.

- *Concurrent validation second-tour (CVII) sample*: The same teams that administered the first-tour performance measures to the LVI sample administered second-tour performance measures at the same location and during the same time periods to a sample of junior NCOs from the 83/84 cohort who were in their second tour of duty (4–5 years of service).
- *Longitudinal validation second-tour (LVII) sample*: This sample included members of the 86/87 cohort from the Batch A MOS who were part of the LVP (predictors), LVT (training performance measures), and LVI (first-tour job performance measures) samples and who reenlisted for a second tour (6–7 years of total service). The second-tour (NCO) performance measures were administered at 15 U.S. posts, multiple locations in Germany, and two locations in Korea.

RESEARCH INSTRUMENT DEVELOPMENT: PREDICTORS

A major objective was to develop an experimental battery of new tests that had maximum potential for enhancing selection and classification decisions for the entire enlisted personnel system. Consequently, rather than basing the selection of predictor constructs on job analyses of the specific occupations in question, although we subsequently did them for purposes of criterion development, the general strategy was to identify a universe of potential predictor constructs for the population of enlisted MOS and then sample appropriately from it. The appropriate constructs were those that were judged to maximize the expected linkages with the population of performance components, not just the performance components encompassed by the 21 MOS in the sample. The next steps were to construct measures for each sampled construct that was above a threshold of criticality, given the measurement purposes. Use of available commercial measures was not considered given the requirements of the military enlistment testing system, although some such tests (e.g., subtests from the Employee Aptitude Survey, Differential Aptitude Test) were used as marker measures in the development of the new measures. Accordingly, the intent was to develop a predictor battery that was maximally useful for selection and classification into an entire population of jobs, and that provided maximal incremental information beyond that provided by the ASVAB.

Critical Constructs

After a thorough literature review (including available meta-analyses), the research team identified a list of 53 potentially useful predictor variables. A sample of 35 personnel selection experts was then asked to estimate the expected correlations between each predictor construct and an array of potential performance factors. All of the information was then used to arrive at a final set of variables ($k = 25$) for which new measures would be constructed.

The measure development process included (a) a large-scale field test that also administered a set of established marker tests for several of the constructs (see Peterson et al., 1990); (b) development of software for a computerized battery of perceptual/psychomotor tests, as well as a portable testing station and a special response pedestal permitting various responses (e.g., one-hand tracking, two-hand coordination); (c) several paper-and-pencil cognitive tests; and (d) two inventories. One inventory assessed relevant vocational interests, and the second focused on major dimensions of personality and biographical history. This process resulted in the following experimental predictor battery that was first used in the Concurrent Validation (CVI) in 1985.

Experimental Cognitive Ability Tests: Paper and Pencil

This section presents a description of the tests grouped by the construct they were intended to measure.

Spatial visualization—rotation

Assembling objects test: Select the alternative that depicts the components assembled correctly.

Object rotation test: Is the figure represented the same as the test object, only rotated?

Spatial visualization—scanning

Maze test: Determine which of the four entrances leads to a correct pathway.

Spatial orientation

Orientation test: Rotate the frame to match correct orientation of a picture.

Map test: From directions to only one landmark infer directions to another.

Reasoning test: Given a series of figures, identify the figure that should appear next.

Experimental Cognitive Ability Tests: Computer-Based

All data for these tests were collected using the custom fabricated response pedestal.

Reaction Time (Processing Efficiency)

Simple reaction time 1: Mean decision time (the time between appearance of the stimulus and the removal of the subject's hand from the home button to strike response button) and movement time (total time to strike response button minus decision time) were computed.

Choice reaction time 2: Making correct choices from two alternatives.

Short-term memory

Memory search test: Memory for previous displays of letters or numbers.

Perceptual speed and accuracy

Perceptual speed and accuracy test: Percent correct and mean decision time for comparison of two visual stimuli presented simultaneously (same or different?).

Target identification test: Item shows a target at the top of the screen and three color-labeled stimuli near the bottom. Identify which stimulus represents the same object as the "target." Percent correct and mean decision times are computed.

Psychomotor Precision

Psychomotor precision encompasses two of the ability constructs identified by Fleishman and his associates: control precision and rate control (Fleishman, 1967).

Target tracking test 1: As a target moves at a constant rate along a path consisting of horizontal and vertical lines, a single joystick is used to keep crosshairs centered on the target. The mean distance from the crosshairs to the center of the target, computed several times each second, constitutes overall accuracy.

Target shoot test: A target moves in an unpredictable manner. A joystick is used to move the crosshairs into the center of the target and "fire." Scored on accuracy and "time to fire."

Multilimb coordination

Target tracking test 2: The subject manipulates two sliding resistors to control movement of the crosshairs: one in the horizontal plane and the other in the vertical plane.

Number operations

Number memory test: Successive arithmetic operations appear on the screen until a solution is presented and the subject must indicate whether the solution presented is correct.

Movement judgment

Cannon shoot test: A "cannon" appears and can "fire" a shell, which travels at a constant speed, at a moving target. Subject must fire so that the shell intersects the target.

Personality/Temperament and Biographical Measures

The biographical and temperament/personality variables were incorporated in an inventory titled the Assessment of Background and Life Experiences (ABLE). A list of the specific scales and the composites into which they are grouped are shown in Table 40.2.

Interest (AVOICE) Factors/Scales

The Vocational Interest Career Examination was originally developed by the Air Force. That inventory served as the starting point for the AVOICE (Army Vocational Interest Career Examination). The intent for the AVOICE was to sample content from all six of the constructs identified in Holland's (1966) hexagonal model of interests, as well as to provide coverage of the vocational areas most important in the Army. The 22 scales assessed by the AVOICE, and the composites into which they are grouped, are shown in Table 40.2. The scales can also be sorted into the Holland factors.

Measurement of Outcome Preferences

On the basis of the extensive literature on job outcomes provided by studies of job satisfaction and work motivation, an inventory was developed that asked the respondent to reflect the strength of his or her preferences for certain job outcomes (e.g., rewards) on a seven-point scale. The final form of the inventory was titled the Job Orientation Blank (JOB). Factor analyses of the field test data suggested that the JOB's 29 items could be grouped into six factors, which were grouped into three composites, as shown in Table 40.2.

TABLE 40.2
Experimental Battery: Composite Scores and Constituent Basic Scores

ASVAB Composites (4)	Computer-Administered Test Composites (8)	ABLE Composite (7)	AVOICE Composites (8)
Quantitative	**Psychomotor**	**Achievement orientation**	**Rugged outdoors**
Math knowledge	Target tracking 1 distance	Self-esteem	Combat
Arithmetic reasoning	Target tracking 2 distance	Work orientation	Rugged individualism
	Cannon shoot time score	Energy level	Firearms enthusiast
Technical	Target shoot distance		
Auto shop		**Leadership potential**	**Audiovisual arts**
Mechanical comprehension	**Movement time**	Dominance	Drafting
	Pooled movement time		Audiographics
Electronics information		**Dependability**	Aesthetics
	Perceptual speed	Traditional values	
Speed	Perceptual speed & accuracy (DT)	Conscientiousness	**Interpersonal**
Coding speed		Nondelinquency	Medical services
Number operations	Target identification (DT)		Leadership guidance
		Adjustment	
Verbal	**Basic speed**	Emotional stability	**Skilled/technical**
Word knowledge	Simple reaction time (DT)		Science/chemical
Paragraph comprehension	Choice reaction time (DT)	**Cooperativeness**	Computers
		Cooperativeness	Mathematics
General science	**Perceptual accuracy**		Electronic communication
	Perceptual speed & accuracy (PC)	**Internal control**	
		Internal control	
Spatial Test Composite (1)	Target identification (PC)		**Administrative**
Assembling objects test		**Physical condition**	Clerical/administrative
Object rotation test	**Basic accuracy**	Physical condition	Warehouse/shipping
Maze test	Simple reaction time (PC)		
Orientation test	Choice reaction time (PC)		**Food service**
Map test		**JOB Composites (3)**	Food service professional
Reasoning test	**Number speed and accuracy**		Food service employee
	Number speed (operation DT)	**High job expectations**	
		Pride	**Protective services**
	Number memory (PC)	Job security	Fire protection
		Serving others	Law enforcement
	Short-term memory	Ambition	
	Short-term memory (PC)		**Structural/machines**
	Short-term memory (DT)	**Job routine**	Mechanics
		Routine	Heavy construction
			Electronics
		Job autonomy	Vehicle operator
		Autonomy	

DT, decision time; PC, proportion correct.

Basic Predictor Composite Scores

The ASVAB together with the experimental tests produced a set of 72 scores. This number was too large for validation analyses that take advantage of idiosyncratic sample characteristics (e.g., multiple regression). Therefore, a series of analyses was conducted to determine a smaller set of predictor composite scores that would preserve the heterogeneity of the full set of basic scores to the greatest extent possible. These analyses included exploratory factor analyses and confirmatory factor analyses guided by considerable prior theory and empirical evidence (McHenry, Hough, Toquam, Hanson, & Ashworth, 1990; Peterson et al., 1990). A final set of 31 composites was identified and is shown in Table 40.2.

Collectively, the Experimental Battery and the ASVAB were intended to be a comprehensive and representative sample of predictor measures from the population of individual differences that are relevant for personnel selection and classification, and which can be measured in a standardized fashion at the time of organizational entry.

JOB ANALYSES AND CRITERION DEVELOPMENT

Job Analyses

In contrast to the predictors, virtually all criterion development in Project A/Career Force was based on extensive job analyses, including task descriptions, critical incident analysis, and interviews with Army SMEs. Relevant job manuals and available Army Occupational Survey results were used to enumerate the complete population of major tasks (100–150) for each MOS. The tasks for each MOS were then grouped into clusters and rated for criticality and difficulty by panels of SMEs.

Additional panels of Army SMEs were used in a workshop format to generate approximately 700 to 800 critical incidents of effective and ineffective performance per MOS that were specific to each MOS and approximately 1,100 critical incidents that could apply to any MOS. For the MOS-specific and Army-wide critical incidents, a retranslation procedure was carried out to establish dimensions of performance.

Together, the task descriptions and critical incident analyses of MOS-specific and Army-wide performance were intended to produce a detailed content description of the major components of performance in each MOS and to provide the basis for the development of the performance criterion measures.

The job analysis goals for the second tour included the description of the major differences in technical task content between first and second tour and the description of the leadership/supervision components of the junior NCO position. The task analysis and critical incident steps used for first tour were also used for second tour. In addition, a special 46-item job analysis instrument, the Supervisory Description Questionnaire, was constructed and used to collect item criticality judgments from SMEs. Consequently, the supervisory/leadership tasks judged to be critical for an MOS became part of the population of tasks for that MOS.

Performance Criteria

The general goals of training performance and job performance measurement were to define, or model, the total domain of performance in some reasonable way and then develop reliable and valid measures of each major factor. The general procedure for criterion development followed a basic cycle of a comprehensive literature review, initial instrument construction based on the job analyses previously described, pilot testing, instrument revision, field-testing, and proponent (i.e., management) review. The specific measurement goals were as follows:

1. Develop standardized measures of training achievement for the purpose of determining the relationship between training performance and job performance.
2. Make a state-of-the-art attempt to develop job sample or "hands-on" measures of job task proficiency.
3. Develop written proceduralized knowledge measures of job task proficiency.
4. Develop rating scale measures of performance factors that are common to all first-tour enlisted MOS (Army-wide measures), as well as for factors that are specific to each MOS.

5. Compare hands-on measurement to paper-and-pencil tests and rating measures of proficiency on the same tasks (i.e., a multitrait, multimethod approach).
6. Evaluate existing administrative records as possible indicators of job performance.

Initial Theory

Criterion development efforts were guided by a model that viewed performance as truly multidimensional. For the population of Army entry-level enlisted positions, there were two major types of performance components: (a) those that reflect specific technical tasks or specific job behaviors that are not required for other jobs and (b) components that are defined and measured in the same way for every job (i.e., Army-wide), such as first aid and basic weapons proficiency, contributions to teamwork, continual self-development, support for the norms and customs of the organization, and perseverance in the face of adversity. The working model of total performance with which Project A began viewed performance as multidimensional within these two broad categories.

Training Performance Measures

Because a major program objective was to determine the relationships between training performance and job performance, a comprehensive training achievement test was constructed for each MOS on the basis of matching the previously determined content of the critical job tasks for each MOS to the program of instruction (POI). For the POI content judged to be reflective of critical job tasks, items were written to represent the proceduralized knowledge reflective of how to do a task. After pilot testing, revision, and review, the result was a 150- to 200-item training achievement test for each MOS. Rating scales were also developed for completion by peers and drill instructors at the end-of-training (EOT).

First-Tour (Entry-Level) Performance Measures

For first-tour performance criteria development, the task-based information was used to develop standardized hands-on job samples, paper-and-pencil job knowledge tests, and rating scales for each Batch A MOS. Roughly 30 critical tasks per MOS were covered by the written job knowledge tests and rating scales, and about one-half of those tasks were also tested using a hands-on format. For the hands-on simulations, each examinee passed through a testing station for each of the 15 ± 2 major job tasks and was asked to perform a standardized simulation of the task, using real equipment, if at all possible.

From the critical incident analyses, a modified behaviorally anchored rating scale procedure was used to construct six to nine rating scales for performance factors specific to a particular job and also for ten performance factors that were defined in the same way and relevant for all jobs. The critical incident procedure was also used with workshops of combat veterans to develop rating scales of predicted combat effectiveness because, except in the late stages of the project, soldiers in our samples did not have combat experience. Ratings were gathered from peers and supervisors of first-tour soldiers. Data collection activities included a comprehensive rater training program.

The final category of criterion measures was produced by a search of the Army's administrative records for potential performance measures, which yielded several promising indicators that were collected largely through self-report.

Second-Tour (NCO) Measures

The job analyses of the second-tour jobs indicated considerable overlap between first-and second-tour technical task content, although NCOs were expected to perform at somewhat higher levels. Consequently, although there were some differences in technical tasks selected for testing, the first-tour technical performance measures were generally used to measure second-tour performance as well. The more substantive differences occur because during their second tour, soldiers begin to take on substantial and critical leadership responsibilities.

The second-tour job analysis results identified six additional MOS-specific leadership dimensions and three Army-wide leadership dimensions. A set of supervisory performance rating scales was created to measure the following dimensions: acting as a role model, communication, personal counseling, monitoring subordinate performance, organizing missions/operations, personnel administration, and performance counseling/correcting. Because it proved infeasible to collect peer ratings from second-tour soldiers in CVII, only supervisor ratings were collected in LVII.

On the basis of a review of the literature and consideration of feasibility, two additional methods were developed for assessing NCO performance. The first was a set of assessment center-like role-play exercises, and the second was a written situational judgment test.

Supervisory Role-Play Exercises

Role-play exercises were developed to simulate three of the critical and distinct NCO supervisory tasks.

1. Counseling a subordinate with personal problems that affect performance
2. Counseling a subordinate with a disciplinary problem
3. Conducting one-on-one remedial training

The format for the simulations was for the examinee to play the role of a supervisor. A trained confederate, who also scored the performance of the examinee on several specific dimensions within the three major tasks, played the subordinate.

Situational Judgment Test

The situational judgment test (SJT) measurement's purpose was to evaluate the effectiveness of judgments about how to react in typical supervisory problem situations. A critical incident methodology was used to generate the situations, and response options were developed through input from senior NCO SMEs and from examinees during the field tests. Independent groups of SMEs scaled the response options in terms of effectiveness, and examinees selected the options they believed would be most and least effective.

MODELING THE LATENT STRUCTURE OF PERFORMANCE

As detailed above, there were three distinct performance domains—training performance, first-tour job performance, and second-tour job performance—and there were many more individual scores for each person than there were on the predictor side (e.g., 150+ scores for first-tour performance). Depending on the instrument, either expert judgment or exploratory factor analysis/cluster analysis was used to identify "basic" composite scores that reduced the number of specific individual scores but attempted to minimize the loss of information. These analyses resulted in 24–28 basic criterion scores for each job, which was still too many for validation purposes.

The next step used all available expert judgment to postulate a set of alternative a priori factor models of the latent structure underlying the covariances among the basic scores. These alternative models were then subjected to a confirmatory analysis using LISREL. The first confirmatory test used the covariance matrix estimated on the CVI sample to evaluate the relative accuracy of fit of the alternative models proposed by the psychologist SMEs. The model that best fit the concurrent sample data was evaluated again by fitting it to the LVI sample data. This was a true confirmatory test. The same procedure was used to determine the best fitting model in the longitudinal sample (LVI), from among a new set of a priori alternatives proposed by SMEs, and then evaluate it again on the CVI data. A similar kind of double cross-validation procedure was followed to test the fit of alternative factor models to the basic criterion score covariances estimated from the concurrent and longitudinal second-tour samples (CVII and LVII). For the first- (entry-level enlisted) and second-tour (junior NCOs), the best fitting models determined independently in the concurrent and longitudinal sample were virtually identical. Also, the best fitting model in one sample (i.e., current or longitudinal) fit the data equally as well in the other sample.

Because there were far fewer criterion measures, a similar confirmatory procedure was not used to model training performance. Instead, expert judgment was used to group the training performance criteria into composites that paralleled the latent factors in the first-tour and second-tour performance models. The expert judgment based factors were then checked against the criterion intercorrelation matrix estimated from the training validation (LVT) sample. The prescribed model fit the data better than any alternative.

In summary, the modeling analyses produced a specification of the factor scores (i.e., latent variables) that defined the latent structure of performance at each of three organizational levels, or career stages: EOT performance, first-tour job performance, and second-tour job performance. It was the performance factor scores at each of these three points that constituted the criterion scores for all subsequent validation analyses.

A Model of Training Performance

As noted previously, the EOT performance measures were intended to parallel the Army-wide rating scales and job knowledge tests used for first-term job incumbents in the same MOS. Consequently, the training achievement tests constructed for each MOS covered Army-wide basic training content and MOS-specific technical content. Of the ten Army-wide first-term rating scales, three had no EOT counterpart. The remaining seven were grouped into clusters to parallel the first-tour ratings factors. A leadership potential scale was added. The six scores obtained from the EOT measures are shown in Figure 40.2.

A Model of First-Tour Job Performance

The result of the confirmatory factor analyses described earlier was a five-factor model of first-term (entry-level) performance that was very robust across samples. The definition of the factors is provided below and the basic criterion scores that comprise them are shown in Figure 40.3.

1. *Core Technical Proficiency (CTP)*: Represents the proficiency with which the soldier performs the tasks that are "central" to the MOS. These tasks represent the core of the job and are its primary definers. This construct does not include the individual's willingness to perform the task.
2. *General Soldiering Proficiency (GSP)*: In addition to the technical content specific to an MOS, individuals in every MOS also are responsible for being able to perform a variety of Army-wide tasks (e.g., first aid, land navigation).
3. *Effort and Leadership (ELS)*: This construct reflects the degree to which the individual exerts effort over the full range of job tasks, perseveres under adverse or dangerous conditions, and demonstrates leadership and support toward peers.

From the EOT achievement test

1. Technical content score (TECH)
2. Army-wide basic training content total score (BASC)

From the EOT rating scales

3. Technical achievement and effort (ETS)
4. Maintaining personal discipline (MPD)
5. Physical fitness and military bearing (PFB)
6. Leadership potential (LDR)

FIGURE 40.2 A model of training performance and associated scores.

1. Core technical proficiency (CTP) • Hands-on test–MOS specific tasks • Job knowledge test–MOS-specific tasks	1. Core technical proficiency (CTP) • Hands-on test–MOS specific tasks • Job knowledge test–MOS specific tasks
2. General soldiering proficiency (GSP) • Hands-on test–common tasks • Job knowledge test–common tasks	2. General soldiering proficiency (GSP) • Hands-on test–common tasks • Job knowledge test–common tasks
3. Effort and leadership (ELS) • Admin: number of awards and certificates • Army-wide rating scales: overall effectiveness rating scale • Army-wide rating scales: effort/leadership ratings factor • Average of MOS specific ratings scales	3. Achievement and effort (AE) • Admin: number of awards and certificates • Army-wide rating scales: overall effectiveness rating scale • Army-wide rating scales: technical skill/effort ratings • Average of MOS-specific rating scales • Average of combat performance prediction rating scales
4. Maintaining personal discipline (MPD) • Admin: number of disciplinary actions • Admin: promotion rate score • Army-wide rating scales: personal discipline ratings factor	4. Maintaining personal discipline (MPD) • Admin: number of disciplinary actions • Army-wide rating scales: personal discipline ratings factor
5. Physical fitness and military bearing (PFB) • Admin: physical readiness test score • Army-wide rating scales: physical fitness/bearing ratings factor	5. Physical fitness and military bearing (PFB) • Admin: physical readiness test score • Army-wide ratings scales: physical fitness/bearing ratings factor
	6. Leadership (LDR) • Admin: promotion rate score • Army-wide rating scales: leading/supervising ratings factor • Individual scores from each of the three role plays • Situational judgment test–total score

FIGURE 40.3 Models of first- and second-tour job performance and associated scores.

4. *Maintaining Personal Discipline (MPD)*: MPD reflects the degree to which the individual adheres to Army regulations and traditions, exercises personal self-control, demonstrates integrity in day-to-day behavior, and does not create disciplinary problems.
5. *Physical Fitness and Military Bearing (PFB)*: PFB represents the degree to which the individual maintains an appropriate military appearance and stays in good physical condition.

Note that the first two factors are represented by the hands-on work sample test and the job knowledge tests, whereas the last three factors are each represented by rating scales and administrative measures. Again, this factor solution represented the best fitting a priori model in the concurrent and longitudinal cohort samples and was cross-validated from one cohort to the other with no loss in the accuracy of fit. It was a very stable representation. It also led to further developments in performance modeling, most notably the eight-factor model described by J. P. Campbell and his colleagues (e.g., Campbell, McCloy, Oppler, & Sager, 1993) and the core technical versus contextual performance model proposed by Borman and Motowidlo (1993).

A Model of Second-Tour Performance

A confirmatory factor analysis procedure similar to that used for the first-tour analysis yielded a six-factor model of NCO performance. The sixth factor represents the leadership supervisory component. The other five factors are very similar in content to the first-term model, as shown in Figure 40.3, which presents the factors and associated basic criterion scores for the first- and second-term models side-by-side. Elements of the two models that differ are identified in italics.

CORRELATIONS OF PAST PERFORMANCE WITH FUTURE PERFORMANCE

The longitudinal component of the Project A design provided an opportunity to collect performance data on the same people at three points in time: (a) at the end of training (LVT), (b) during the first tour of duty (LVI), and (c) during the second tour of duty (LVII). It was virtually an unparalleled opportunity to examine the consistencies in performance over time from the vantage point of multiple jobs, multiple measures, and a substantive model of performance itself.

This question encompasses at least two specific issues. First, the degree to which individual differences in future performance can be predicted from individual differences in past performance is a function of the relative stability of performance across time. Do the true scores for individuals change at different rates even when all individuals are operating under the same "treatment" conditions? The arguments over this question sometimes become a bit heated (Ackerman, 1989; Austin, Humphreys, & Hulin, 1989; Barrett & Alexander, 1989; Barrett, Caldwell, & Alexander, 1985; Henry & Hulin, 1987; Hulin, Henry, & Noon, 1990).

The second issue concerns whether the current and future jobs possess enough communality in their knowledge, skill, or other attribute requirements to produce valid predictions of future performance from past performance. Perhaps the determinants of performance on the new job are simply too different. For example, the degree to which "managers" should possess domain-specific expertise has long been argued. Just as an army should not be equipped and trained to fight only the last war, the promotion system should not try to maximize performance in the previous job. The data from Project A permit some of the above issues to be addressed, and the models of performance for training, first-tour performance, and second-tour provide some clear predictions about the pattern of convergent and divergent relationships.

The LVT × LVI, LVI × LVII, and LVT × LVII intercorrelations, corrected and uncorrected for attenuation, are reported in Reynolds, Bayless, and Campbell (2001). Only the correlations between first- and second-tour performance are shown here (Table 40.3). Three correlations are shown for each relationship. The top figure is the mean correlation across MOS corrected for restriction of range (using the training sample as the population) but not for attenuation. The first value in the parentheses is this same correlation after correction for unreliability in the measure of "future" performance, or the criterion variable when the context is the prediction of future performance from past performance. The second value within the parentheses is the value of the mean intercorrelation after correction for unreliability in the measure of "current" performance and the measure of future performance. The reliability estimates used to correct the upper value were the median values of the individual MOS reliabilities.

The pattern of correlations in Table 40.3 exhibits considerable convergent and divergent properties. The most interesting exception concerns the prediction of second-tour leadership performance. Virtually all components of previous performance are predictive of future leadership performance, which has important implications for modeling the determinants of leadership. For example, on the basis of the evidence in Table 40.3, one might infer that effective leadership is a function of being a high scorer on virtually all facets of performance. The least critical determinant is military bearing and physical fitness, which some might call "looking like a leader." Project A provides the only existing data set for examining these issues. A surprisingly similar pattern of relationships was found when training performance was used as a predictor of first-tour performance and of NCO performance. The average corrected correlation between the rating of leadership potential for trainees and rated leadership potential for first-term individuals was .58. In general, the results were consistent, meaningful, and stronger than anyone expected.

TABLE 40.3
Zero-Order Correlations of First-Tour Job Performance (LVI) Variables With Second-Tour Job Performance (LVII) Variables: Weighted Average Across MOS

	LVI:CTP	LVI:GSP	LVI:ELS	LVI:MPD	LVI:PFB	LVI:NCOP
LVII: Core technical proficiency (CTP)	**.44** (.55/.59)	.41 (.49/.55)	.25 (.30/.33)	.08 (.10/.11)	.02 (.02/.03)	.22 (.26/.29)
LVII: General soldiering proficiency (GSP)	.51 (.60/.68)	**.57** (.67/.76)	.22 (.26/.29)	.09 (.11/.12)	−.01 (−.01/−.01)	.19 (.22/.25)
LVII: Effort and achievement (EA)	.10 (.11/.12)	.17 (.18/.20)	**.45** (.49/.53)	.28 (.30/.33)	.32 (.35/.38)	.43 (.46/.50)
LVII: Leadership (LEAD)	.36 (.39/.42)	.41 (.44/.47)	**.38** (.41/.45)	.27 (.29/.32)	.17 (.18/.20)	.41 (.44/.48)
LVII: Maintain personal discipline (MPD)	−.04 (−.04/−.05)	.04 (.04/.05)	.12 (.13/.15)	**.26** (.29/.32)	.17 (.19/.21)	.16 (.18/.20)
LVII: Physical fitness and bearing (PFB)	−.03 (−.03/−.04)	−.01 (−.01/−.01)	.22 (.24/.27)	.14 (.15/.17)	**.46** (.51/.56)	.30 (.33/.36)
LVII: Rating of overall effectiveness (EFFR)	.11 (.14/.16)	.15 (.19/.22)	.35 (.45/.49)	.25 (.32/.36)	.31 (.40/.44)	.41 (.53/.68)

Total pairwise *N* values range from 333 to 413. Correlations corrected for range restriction. Correlations between matching variables are in **bold**. Leftmost coefficients in parentheses are corrected for attenuation in the future criterion. Rightmost coefficients in parentheses are corrected for attenuation in both criteria. ELS, effort and leadership; NCOP, NCO potential—a single scale.

CRITERION-RELATED VALIDATION

Types of Information

Project A/Career Force produced a very large archive of validation results. The following is a basic list of the major categories of information.

1. For the CVI and LVI samples (i.e., the prediction of first-tour performance), the basic data set consisted of the zero-order correlations of each basic predictor score with each of the five performance factors. This intercorrelation matrix was computed for each MOS in each of the CVI and LVI samples. Then for each sample, the full prediction equation (e.g., each predictor variable was included) was evaluated for each criterion factor, using several different kinds of predictor weights. For comparative purposes, the estimates were corrected for restriction of range and for criterion unreliability.
2. The incremental validity of each predictor domain over ASVAB was evaluated for predicting each of the five performance factors in CVI and LVI.
3. The same basic validities and the incremental validities were estimated for the CVII and LVII samples, except that the six factors in the second-tour (NCO) performance model were used as criteria.
4. For the LVII sample, the ASVAB, the Experimental Battery, and assessment of performance during the first tour were weighted via hierarchical regression to estimate the incremental validities over ASVAB produced by adding the Experimental Battery first and then the first-tour performance assessment to the prediction of performance in the second tour.

5. The LVI sample data were used to estimate the overall classification gains (compared to gains from selection only) when particular predictor batteries were used to make optimal assignments to MOS.
6. Using a nine-factor measure of the perceived leadership environment for enlisted personnel that was administered during CVI, the moderator effects of the perceived leadership environment on the relationship between cognitive ability and first-tour performance and the relationship between personality and first-tour performance were estimated.

Selected Highlights of Results

The Project A/Career Force data archive has been used in literally hundreds of different analyses pertaining to many kinds of research questions. Only a few of the highlights are discussed here. For more detail, the reader should consult the special summer issue of *Personnel Psychology* (Campbell, 1990) and the Project A "book" (Campbell & Knapp, 2001), as well as the many journal articles and technical reports referenced in those sources.

ASVAB validities were estimated twice for each major factor of first-tour performance and twice for each major factor of second-tour performance. As shown in Table 40.4, the profiles of validity estimates (i.e., across performance factors) were very similar for each of the samples (e.g., .62 to .65, on the average across MOS, for predicting the Core Technical Proficiency factor). Correcting for unreliability in the criterion pushes the estimates (not shown) close to .70. ASVAB predicts job performance in the Army as well as it does training performance, and the estimated validities are quite high.

In general, ASVAB tends to be the best predictor of each of the performance factors in each of the major data sets, although the spatial tests, the computer-based cognitive tests, and the personality and interest measures have substantial correlations as well. The personality scales tend to have slightly higher correlations with the personal discipline factors. The relatively high correlation of the interest scales with the technical performance factors and with the leadership related factors was somewhat unexpected, given the extant literature.

Incremental validities were primarily concentrated in the prediction of the peer leadership and personal discipline factors by the ABLE scales. At the MOS level, specific psychomotor tests

TABLE 40.4
Comparison of Multiple Correlations Averaged Across Batch A MOS When the ASVAB Factors, the Spatial, Cognitive Computer-Based Scores, ABLE Composites, and AVOICE Composites Are Used as Predictors

	ASVAB Factors			Spatial Composite			Computer-Based Scores			ABLE-Based Scores			AVOICE Basic Scores		
Criterion	**LV**	**CV**	**LVII**	**LV**	**CV**	**LVII**	**LV**	**CV**	**LVII**	**LV**	**CV**	**LVII**	**LV**	**CV**	**LVII**
CTP	63	63	64	57	55	57	50	53	53	27	26	24	41	35	41
GSP	67	65	63	64	63	58	57	57	48	29	25	19	40	34	29
ELS (AE)	39	31	29	32	25	27	34	26	09	20	33	13	25	24	09
MPD	22	16	15	14	12	15	15	12	12	22	32	06	11	13	06
PFB	21	20	16	10	10	13	17	11	03	31	37	17	15	12	09
LDR			63			55			49			34			35

Results corrected for range restriction and adjusted for shrinkage. Decimals omitted. CTP, core technical proficiency; CV, concurrent validation; ELS (AE), effort and leadership (achievement and effort); GSP, general soldiering proficiency; LDR, leadership; LV, longitudinal validation (first-term); LVII, longitudinal validation (second-term); MPD, maintain personal discipline; PFB, physical fitness and bearing.

incremented ASVAB for core technical components that involved firing hand operated weapons (e.g., TOW missile) (Walker & Rumsey, 2001).

An estimate of the maximum validity attainable from ASVAB plus the Experimental Battery for predicting Core Technical Proficiency is shown in Table 40.5. The "reduced equation" is comprised of the four ASVAB factors plus six to eight scores (depending on the MOS) from the Experimental Battery that were chosen a priori by the psychologist SMEs as the most likely to increment ASVAB. The results are shown for unit weights, zero-order validity coefficients as weights, and multiple regression weights with the *R* values adjusted for shrinkage and corrected for unreliability in Core Technical Proficiency.

For the prediction of second-tour performance, as shown in Table 40.4, the estimated validities of the cognitive ability tests for predicting Core Technical Proficiency and General Soldiering Proficiency were virtually identical for first tour (2–3 years after enlistment) and for second tour (6–7 years after enlistment). Overall, the validities did not degrade, as some have speculated they should (e.g., Henry & Hulin, 1987). The only real change was for predicting Effort and Leadership/Effort and Achievement, but the nature of this performance dimension in fact changed between first tour and second tour. For second-tour, the leadership components formed a separate factor.

As noted previously, the first-tour performance factors had substantial correlations with the second-tour performance factors and displayed considerable convergent and divergent validity. The assessment of first-tour performance (i.e., in LVI) also provided considerable incremental validity over ASVAB and the Experimental Battery for predicting the Effort/Achievement and Leadership performance factors in the second tour, but not for the Core Technical Proficiency factor, as shown in Table 40.6. This reflects, in part, the increased importance of these factors for the NCO position.

Estimating potential classification gains from new predictors is a complex issue that depends on a number of contextual parameters (e.g., quotas, assignment priorities, selection ratio, differential recruitment costs across MOS, and variability in selection validities across jobs). The Project A database for LVI was used to model a multivariate normal population, and Monte Carlo procedures were then used to generate estimates of potential classification gains from using various forms of the Experimental Battery plus ASVAB. Quotas for the nine MOS were set proportional to 1993 accessions.

There are two cross-validation issues in estimating classification gains. One concerns the weights used for the predictor equation for each MOS (e.g., ordinary least-squares weights are sample specific to some degree), and the second concerns the sample specificity of the differential assignments themselves. During Project A/Career Force, a new index of classification efficiency labeled "mean average performance" (MAP) was developed and Monte Carlo methods were used to provide an unbiased estimate of the gains in aggregate performance resulting from classification, as compared with selection alone (Rosse, Campbell, & Peterson, 2001). Using a test battery for each MOS that

TABLE 40.5
Comparison of LVI Estimates of Maximum Predictive Validity, Averaged Over Batch A MOS, When Unit Weights, Zero-Order Validities, or Multiple Regression Weights Are Used to Weight Predictors (Criterion Is Core Technical Proficiency)

Comparison	Unit Weights	Validity Weights	Adjusted *R*	Corrected *R*[a]
Full equation (all predictors)	.57	.70	.70	(.77)
Reduced equation (for selection)	.67	.72	.72	(.79)

All estimates corrected for restriction of range.

[a] Corrected for criterion unreliability.

TABLE 40.6
Multiple Correlations for Predicting Second-Tour Job Performance (LVII) Criteria From ASVAB and Various Combinations of ASVAB, Selected Experimental Battery Predictors, and First-Tour (LVI) Performance Measures: Corrected for Restriction of Range and Criterion Unreliability

	Predictor Composite			
LVII Criterion	**Type of Estimate**	***A***	***A + X***	***A + X + 1***
Core technical proficiency	Adjusted *R*	64	69	68
	Unit weight	52	39	42
Effort/achievement	Adjusted *R*	00	00	38
	Unit weight	16	13	21
Leadership	Adjusted *R*	36	43	65
	Unit weight	40	43	50

Adjusted *R* values from Rozeboom (1978; formula 8). Decimals omitted. A, ASVAB factors (quantitative, speed, technical, verbal). X, experimental battery (spatial, rugged/outdoors interests from AVOICE, achievement orientation, adjustment, physical condition, internal control, cooperativeness, dependability, and leadership from ABLE). 1, the LVI "can do" composite (CTP + GSP) for CTP and "will do" composite (ELS + MPD + PFB) for effort/achievement and leadership. See Table 40.3 for LVI performance factor labels.

was selected, on a priori grounds, to be potentially useful for classification purposes, the aggregate gain in MAP for Core Technical Proficiency was .14 standard deviation (SD) units if all accessions must be classified, and .22 SD units if 5% could remain unassigned. In an organization the size of the Army, such gains would have enormous utility.

Parallel to Project A, ARI sponsored Project B, which developed the Enlisted Personnel Assignment System (EPAS) and uses linear programming strategies to make optimal MOS assignments that maximize a specific function (aggregate performance, training cost savings, number of individuals above a performance minimum, etc.) given a set of constraints to be accommodated (MOS quotas, priorities, selection ratios, etc.). Together, the Project A database and the EPAS algorithm provided an unparalleled test bed for estimating the effects of various selection and classification strategies (Konieczny, Brown, Hutton, & Stewart, 1990).

In addition to the above, many additional data collections and analyses were carried out pertaining to such issues as (a) estimating the differential utility of performance gains across jobs, (b) the differential criticality of specific performance factors across jobs, (c) the prediction of attrition, (d) the influence of reward preferences on performance prediction, and (e) the influence of race and gender on performance assessments. The reader must consult Campbell and Knapp (2001) and the articles and technical reports they referenced for the details.

SOME BROADER IMPLICATIONS

It is all well and good that the project did what it proposed to do on time, that it was a rewarding experience for the participants, and that it provided massive evidence for the validity of ASVAB, but what are the broader implications of its substantive outcomes? We list a critical few.

Job and Occupational Analysis

For purposes of developing measures of individual performance, the strong conclusion must be that one method of job analysis is not enough. Different methods (e.g. task analysis, critical incidents) give somewhat different, but complementary, pictures of performance requirements. There is probably no personnel research purpose that would not be better served by multiple methods.

Importance of Taxonomic Theory

The necessity of thinking in terms of the latent structure soon became apparent to everyone. Even the diehards were pushed in this direction because the specific MOS in the sample were not the primary interest.

Implications for Performance Measurement

Project A presented the first real opportunity to investigate the latent structure of performance in this way. We believe it helped change the way industrial-organizational psychologists think about the "criterion problem" (or at least how they should think about it) and about performance measurement in general, regardless of the purpose (Knapp, 2006).

Rating Method

The rating method represents a complex process of information processing and social cognition that is rampant with opportunities for biased and error-filled judgments (e.g., Morgeson & Campion, 1997). It has a bad press. However, Project A rating measures yielded reasonable distributional properties, had reasonable single-rater reliabilities across cohorts, and produced a factor structure that was highly replicable (virtually to the second decimal place). One conclusion might be that ratings are a valuable measurement method if (a) the dimensions to be rated are carefully defined and meaningful to the raters, (b) there is considerable rater training, (c) the setting ensures that raters will give sufficient time and attention to the rating task, and (d) the goals of the rater are commensurate with the goals of the researchers. That is, the measurement is for research purposes, not operational performance appraisal, and the raters accept the goal of doing their best to assess performance on the dimensions as defined by the researchers. Although minimal, such conditions probably go far beyond most studies.

Role of the Measurement Goal

A frequently asked question concerns which type of criterion measure is "best," which implies there must be a near-ultimate criterion lurking someplace. For example, the National Research Council panel (Wigdor & Green, 1991) took the position (we think in error) that the hands-on job sample simulation is the preferred criterion measure, always. The intent of Project A was to counter the argument that there is always one preferred measurement method. Different measurement methods permit different sources of variation to operate (McCloy, Campbell, & Cudeck, 1994), and the choice of measurement method depends on the potential sources of variation that the investigator or practitioner wants to capture, not on being more or less ultimate.

General Factor Versus Specific Factors

Because of the generally positive manifold in the intercorrelation matrix for any set of job performance measures, even when method variance and unreliability are controlled (Viswesvaran, Schmidt, & Ones, 1993), there will always be a general factor. However, the general factor is not there because of only one general performance requirement. It arises most likely because individual differences in general mental ability (GMA) and individual differences in the predisposition toward conscientious and effort are determinants of performance on virtually all aspects of most jobs, even for performance requirements that entail very different content (e.g., electronic troubleshooting vs. rewarding subordinates appropriately). However, a general factor does not preclude either the existence or the importance of specific factors for selection and classification. The naïve use of the term "overall performance" should become a thing of the past. There is no substantive construct that can be labeled as general, or overall, performance. Anyone who tries to provide specifications for such a construct simply cannot do it. They must resort to combining the verbal specifications for the individual specific factors. Now, anyone can add up the "scores" on the specific factors to obtain a total score; and a weighted sum may indeed be necessary for certain specific decision purposes (Schmidt & Kaplan, 1971). MacKenzie, Podsakoff, and Jarvis (2005) refered to such a score as a formative score in contrast to a reflective score, which is intended to represent a substantive latent variable. For confirmatory and reliability estimation purposes, the two require a different measurement model (see Chapter 2, this volume), and Mackenzie et al. (2005) discussed how model misspecifications can lead to faulty inference.

Role of Personality

In retrospect, the development and validation of the ABLE was one of the primary reasons for the resurgence of research on personality for selection purposes, along with the subsequent meta-analysis reported by Barrick and Mount (1991) that covered nonmilitary occupations. However, the ABLE results, which differed in substantive ways between the concurrent and longitudinal validations, also introduce cautions about the reactivity to experience of some types of items and the subsequent use of concurrent designs for validating personality measures.

GMA Versus Differential Prediction

There is certainly no denying the dominant role of GMA in the prediction equation for virtually any job. However, the principal lessons from Project A are that the degree to which incremental validity and/or differential validity are possible is influenced significantly by the component of performance being predicted and the range of predictor variables that can be used.

Estimating Classification Efficiency

The Project A database also provided a rare opportunity to estimate classification gains under a variety of conditions, without having to assume the simplifying conditions required by the Brogden-type estimator. Zeidner and Johnson and their colleagues (e.g., Scholarios, Johnson, & Zeidner, 1994; Zeidner, Johnson, & Scholarios, 1997) carried out an extensive series of Monte Carlo simulation studies using the Project A database and showed that small but operationally significant classification gains could be realized using only a battery of ability tests. Our own analyses showed that larger estimated gains could be obtained if the entire Project A Experimental Battery, plus ASVAB, could be used.

Beyond the Army

The original objectives set by the sponsor were met. But what about the Army as a unique organization and the generalization of any findings to the civilian sector? Certainly in some respects the Army is a unique organization. No civilian organization has a similar mission, and some of the components of individual performance have unique aspects. However, the bulk of the performance domain for the enlisted occupational structure has civilian counterparts and the personnel corps is reasonably representative of the civilian labor force in similar jobs with similar levels of experience. It is our firm belief that the major implications of the project's methods and results are not constrained by the uniqueness of the Army as an organization and have broad applicability to understanding the world of work.

CONCLUSIONS

The experiences of Project A argue again and again for the necessity of placing the measures and variables used in a particular study within a model of the relevant latent structure that is specified as well as possible. It facilitates the interpretation of results, the integration of findings across studies, the identification of future research needs, the use of the findings for unanticipated applications, and the generalization of findings to other settings.

Continually trying to improve our models of relevant domains, as well as the interrelationship among them, is critical for (good) practice as it is for (good) science. We began Project A with high respect for our practice and our science. The respect for both and the appreciation for how they are so strongly interrelated were even greater at the end.

REFERENCES

Ackerman, P. L. (1989). Within task intercorrelations of skilled performance: Implications for predicting individual differences? (A commentary on Henry & Hulin, 1987). *Journal of Applied Psychology, 97*, 360–364.

Austin, J. T., Humphreys, L. G., & Hulin, C. L. (1989). A critical reanalysis of Barrett et al. *Personnel Psychology, 42*, 583–596.

Barrett, G. V., & Alexander, R. A. (1989). Rejoinder to Austin, Humphreys, and Hulin: Critical reanalysis of Barrett, Caldwell, and Alexander. *Personnel Psychology, 42*, 597–612.

Barrett, G. V., Caldwell, M. S., & Alexander, R. A. (1985). The concept of dynamic criteria: A crucial reanalysis. *Personnel Psychology, 38*, 41–56.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance. A meta-analysis. *Personnel Psychology*, *44*, 1–26.

Borman, W. C., & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt, W. C. Borman, & Associates (Eds.), *Personnel selection in organizations* (pp. 71–98). San Francisco, CA: Jossey-Bass.

Brogden, H. E. (1946). An approach to the problem of differential prediction. *Psychometrika, 11*, 139–154.

Campbell, J. P. (1990). An overview of the Army selection and classification project (Project A). *Personnel Psychology, 43*, 231–239.

Campbell, J. P., & Knapp, D. J. (2001). *Exploring the limits in personnel selection and classification*. Mahwah, NJ: Lawrence Erlbaum.

Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt & W. C. Borman (Eds.), *Frontiers in industrial/organizational psychology: Personnel selection and classification* (pp. 35–71). San Francisco, CA: Jossey-Bass.

Fleishman, E. A. (1967). Performance assessment based on an empirically derived task taxonomy. *Human Factors, 9,* 349–366.

Ghiselli, E. E. (1973). The validity of aptitude tests in personnel selection. *Personnel Psychology, 26*, 461–477.

Henry, R. A., & Hulin, C. L. (1987). Stability of skilled performance across time: Some generalizations and limitations on utilities. *Journal of Applied Psychology, 72*, 457–462.

Holland, J. L. (1966). *The psychology of vocational choice.* Waltham, MA: Blaisdell.

Hulin, C. L., Henry, R. A., & Noon, S. L. (1990). Adding a dimension: Time as a factor in the generalizability of predictive relationships. *Psychological Bulletin, 107*, 328–340.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 98*, 72–98.

Knapp, D. J. (2006). The U.S. Joint-Service Job Performance Measurement Project. In W. Bennett, C. E. Lance, & D. J. Woehr (Eds.), *Performance measurement: Current perspectives and future challenges* (pp. 113–140). Mahwah, NJ: Lawrence Erlbaum.

Konieczny, F. B., Brown, G. N., Hutton, J., & Stewart, J. E. (1990). *Enlisted Personnel Allocation System: Final report* (ARI Technical Report 902). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Mackenzie, S. B., Podsakoff, P. M., & Jarvis, C. D. (2005). The problem of measurement model misspecification in behavioral and organizational research and some recommended solutions. *Journal of Applied Psychology, 90*, 710–730.

McCloy, R. A., Campbell, J. P., & Cudeck, R. (1994). A confirmatory test of a model of performance determinants. *Journal of Applied Psychology, 79*, 493–504.

McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. A., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. *Personnel Psychology, 43*, 335–354.

Morgeson, F. P., & Campion, M. A. (1997). Social and cognitive sources potential of inaccuracy in job analysis. *Journal of Applied Psychology, 82*, 627–655.

Nathan, B. R., & Alexander, R. A. (1988). A comparison of criteria for test validation: A meta-analytic investigation. *Personnel Psychology, 41*, 517–536.

Peterson, N. G., Hough, L. M., Dunnette, M. D., Rosse, R. L., Houston, J. S., Toquam, J. L., & Wing, H. (1990). Project A: Specification of the predictor domain and development of new selection/classification tests. *Personnel Psychology, 43*, 247–276.

Reynolds, D. H., Bayless, A., & Campbell, J. P. (2001). Criterion reliability and the prediction of future performance from prior performance. In J. P. Campbell & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification*. Mahwah, NJ: Lawrence Erlbaum.

Rosse, R. L., Campbell, J. P., & Peterson, N. G. (2001). Personnel classification and differential job assignments: Estimating classification gains. In J. P. Campbell & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification*. Mahwah, NJ: Lawrence Erlbaum.

Schmidt, F. L. (1988). Validity generalization and the future of criterion related validity. In H. Wainer & H. Braun (Eds.), *Test validity*. Hillsdale, NJ: Lawrence Erlbaum.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274.

Schmidt, F. L., & Kaplan, L. B. (1971). Composite vs. multiple criteria: A review and resolution of the controversy. *Personnel Psychology, 24*, 419–434.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Meta-analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37*, 407–422.

Scholarios, T. M., Johnson, C. D., & Zeidner, J. (1994). Selecting predictors for maximizing the classification efficiency of a battery. *Journal of Applied Psychology, 79*, 412–424.

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (1993, April). *Theoretical implications of a general factor in job performance criteria.* Paper presented at the 8th Annual Conference of the Society of Industrial and Organizational Psychology. San Francisco, CA.

Walker, C. B., & Rumsey, M. G. (2001). Application of findings: ASVAB, new aptitude tests, and personnel classification. In J. P. Campbell & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification*. Mahwah, NJ: Lawrence Erlbaum.

Wigdor, A. K., & Green, B. F. (1991). *Performance assessment for the workplace (Vols. I-II).* Washington, DC: National Academy Press.

Zeidner, J., Johnson, C. D., & Scholarios, D. (1997). Evaluating military selection and classification systems in the multiple job context. *Military Psychology, 9,* 169–186.