SENSITIVITY AND FAIRNESS OF THE ARMED SERVICES VOCATIONAL APTITUDE BATTERY (ASVAB) TECHNICAL COMPOSITES

Lauress Wise
John Welsh
Defense Manpower Data Center

Frances Grafton
Army Research Institute

Paul Foley
Navy Personnel Research and Development Center

James Earles
Linda Sawin
Armstrong Laboratory

D. R. Divgi
Center for Naval Analyses

December 1992

Approved for public release; distribution is unlimited.

Personnel Testing Division
DEFENSE MANPOWER DATA CENTER

TABLE OF CONTENTS

TABLES

FIGURES

EXECUTIVE SUMMARY

The Government Accounting Office (GAO) issued a report, ***Military Training: Its Effectiveness for Technical Specialties is Unknown*** (GAO, 1990), which raised a number of issues about the cognitive tests used in selecting recruits for technical specialties. The GAO noted that scores on the technical subtests of the Armed Services Vocational Aptitude Battery (ASVAB) were lower for minority and female applicants and asked the Office of the Assistant Secretary of Defense (Force Management and Personnel) to initiate research to identify more sensitive predictors of classroom and job performance for female and minority applicants. The Personnel Testing Division (PTD) of the Defense Manpower Data Center (DMDC), as executive agent for the ASVAB Research and Development, was subsequently asked to coordinate the requested investigation.

The attached report, ***Sensitivity*** and ***Fairness of the Armed Services Vocational Aptitude Battery (ASVAB) Technical Composites,*** is the first result of the investigation. This report describes an extensive assessment of the sensitivity and fairness of the current technical composites for females and blacks. The assessment covered a large number of specialties for which technical subtests (Auto and Shop Information, Electronics Information, and Mechanical Comprehension) are used in selection. Table 1 on page 2 lists the individual subtests of the ASVAB, and Table 2 on page 2 lists the selection composites included in the present analyses.

The data analyzed included final school grades (FSG) for Air Force and Navy technical training courses and Skill Qualification Test (SQT) data on first-term recruits for Army specialties. The samples analyzed included a total of 33,017 females, 249,712 males, 95,080 blacks, and 281,063 whites. Marine Corps job-performance measurement data were analyzed separately. (See Appendix A beginning on page 29.)

The basic definition of sensitivity used in these analyses was the slope of the regression line relating training or job outcomes to selection composite scores.

- The predictor was considered ***sensitive*** if differences in predictor scores were associated with significant differences in the outcomes.

- The predictor composites were considered ***fair*** if individuals at the same score level had the same average outcome regardless of race or gender.

A number of technical issues were addressed in the analyses. These included rescaling the different criterion measures onto a common metric, avoiding problems due to the necessity of using selected samples (trainees and job incumbents in comparison to all applicants), determining the most meaningful way to aggregate results across a large number of different samples, and testing for overall significance.

The basic results, aggregated across both specialties and technical composites, are illustrated in Figures 1 and 2 on page 21. The key findings were:

- the composites were highly sensitive for all groups studied;

- the composites were slightly more sensitive for females in comparison to males and for whites in comparison to blacks, but these differences were too small to be of practical significance; and

- prediction lines were quite similar for all groups.

Overall, female and black performance in both training and on-the-job was somewhat lower than the performance of males and whites. Some, but not all, of these differences were explained by differences in the ASVAB composite scores. The findings were quite similar for each of the individual ASVAB composites included in the study.

The results indicate that the current technical composites **are** sensitive and fair for females and blacks. Nonetheless, use of the technical composites does create a significantly greater barrier for these groups in comparison to males and whites.

The next phase of investigation will focus on alternatives to the current predictors. These alternatives will include evaluation of existing **subtests** and may include new measures now being evaluated for inclusion in future ASVAB forms.

University of Illinois at Urbana-Champaign

Department of Psychology
603 East Daniel Street
Champaign, **IL 61820**

217 244-5876 *fax*

College of Liberal Arts and Sciences

September 14, 1992

Dr. W. S. Sellman
Director for Accession Policy
OASD (FM&P) (MM&PP)
Room 2B271; The Pentagon
Washington, DC 20301-4000

Dear Dr. Sellman:

In May 1991, the Department of Defense Advisory Committee on Military Personnel Testing (DAC) was briefed on a report by the General Accounting Office (Military Training: Its Effectiveness for Technical Specialties is Unknown, GAO:PEMD-91-4, October 1990) that raised a number of issues concerning the fairness and effectiveness of the ASVAB tests currently used in selecting applicants for Enlisted technical specialties. The DAC also carefully read the GAO technical report.

Subsequent to the issuance of the GAO report, you directed the Personnel Testing Division (PTD) at the Defense Manpower Data Center (DMDC), as the executive agent for the ASVAB, to follow through on a GAO recommendation that DOD conduct research to "identify more sensitive predictors of classroom performance for women and minority students from the ASVAB data it already possesses." The DAC has been keenly interested in this research and has been briefed several times by PTD as its work has progressed. The DAC has had numerous questions and suggestions, and commends PTD for the thoughtfulness and thoroughness of its responses.

Standard 1.21 from Standards for Educational and Psychological Testing, jointly published by the American Educational Research Association, the American Psychoiogicai Association, and the National Councii on ivleasurement in Education in 1985 states "When studies of differential prediction are conducted, the reports should include regression equations (or an appropriate equivalent) computed separately for each group..." and comments further that "Correlation coefficients provide inadequate evidence for or against a differential prediction hypothesis if groups ... are found not to be approximately equal with respect to both test and criterion variances." Because there are mean differences in scores on ASVAB technical subtests across racial and gender groups and because applicants for enlistment in technical training schools must exceed certain standards to enlist, there are undoubtedly group differences in test score variances. Thus, correlational analysis cannot provide accurate information about the fairness or unfairness of ASVAB subtests.

The DAC has now reviewed a report (Sensitivity and Fairness of ASVAB Technical Composites, Wise et al., 1992) summarizing the research conducted in response to the issues raised by the GAO. The Wise report describes in very careful detail the data sets that were compiled and analyses that were performed. 'the data sets provided by the Services to PTD are very large and allow definitive answers to the concerns expressed by GAO. 'the analyses performed by **PTD** use regression methods and are thus based on the technically correct approach. The conclusions from **PTD's** analyses -- that the ASVAB technical **subtests** are fair and sensitive (as these terms are defined in the Wise report) -- are clear and compelling. The DAC therefore endorses the conclusions of this report, urges wide dissemination of its results, and encourages sharing the data sets used in the PTD analyses with other interested researchers.

As acknowledged in the Wise report, the adverse impact on minorities and females due to their frequent lack of experience with material covered in the technical **subtests** is incontrovertible. The DAC strongly encourages DOD to continue to explore options, particularly those involving changes in training as well as testing, that might remediate current race and gender differences, and make technical jobs more accessible to all groups of applicants.

Cordially,

Fritz Drasgow

Fritz Drasgow
Chair, Defense Advisory
Committee on Military
Personnel Testing

SENSITIVITY AND FAIRNESS OF THE
ARMED SERVICES VOCATIONAL APTITUDE BATTERY
(ASVAB)
TECHNICAL COMPOSITES

Introduction

In an evaluation of the effectiveness of military technical training, the Government Accounting Office (GAO) raised a number of issues concerning the fairness and effectiveness of the tests currently used in selecting applicants for Enlisted technical specialties (GAO, 1990). Among the conclusions listed in the executive summary of the GAO's report were:

> Women and members of minority groups consistently scored lower in testa used to assign recruits to more technical occupational specialties such as radar specialist positions.
>
> GAO concluded that, for most recruits, the services' selection criteria are moderately successful at predicting individual performance during classroom training. However, they are notably less successful for women and minority recruits.
>
> Each service has evaluation mechanisms in place, but only the Army systematically collects data on the field performance of individual graduates in a way that would allow comparison of a graduate's on-the-job performance with his or her entry-level ability and classroom performance. These data reveal an even weaker connection for women and minority group members between criteria used to assign them to technical specialties and their later field performance....
>
> GAO concluded that the insensitivity of selection and placement measures as predictors of future success for women and minority recruits is a matter of serious concern in view of the military's increasing reliance on these groups to perform technical roles (p. 3).

Subsequent to the issuance of this report, the Director of Department of Defense Accession Policy asked the Defense Manpower Data Center (DMDC), as executive agent for the Armed Services Vocational Aptitude Battery (ASVAB), to prepare a response to the GAO's recommendation that DoD conduct research to "identify more sensitive predictors of classroom performance for female and minority students from the ASVAB data it already possesses" (p. 54). This report describes the results of efforts conducted with the Services to respond fully to the GAO's recommendation.

Background

The fact that scores on the ASVAB technical subtests are, on average, lower for females and minorities is well known on the basis of results from the 1980 norming study. (See Eitelberg, 1988, for a recent analysis of race and gender differences in the ASVAB subtest and composite scores.) However, concerns that the technical subtests may be less sensitive predictors of success in technical training and success in performing technical jobs are new and have not been well studied. Prior research has generally supported the fairness of the ASVAB for both minorities and females. A brief summary of that research is provided here as background for the present study. Table 1 lists the individual subtests of the ASVAB, and Table 2 lists the selection composites included in the present analyses.

Table 1
Current ASVAB Content (Forms 8-22)

	Subtest	Number of Items	Time in Minutes
1.	General Science (GS)	25	11
2.	Arithmetic Reasoning (AR)	30	36
3.	Word Knowledge (WK)	35	11
4.	Paragraph Comprehension (PC)	15	13
5.	Numercial Operations (NO)	50	3
6.	Coding Speed (CS)	84	7
7.	Auto & Shop Information (AS)	25	24
8.	Mathematics Knowledge (MK)	25	11
9.	Mechanical Comprehension (MC)	25	19
10.	Electronics Information (EI)	20	9
	Total	334	144
11.	Verbal Ability (VE) = WK + PC		

Table 2
Current Service Technical Composites

Code	Composite Name	Definition
	AIR FORCE	
M	Mechanical	MC + GS + 2AS
E	Electronics	AR + MK + EI + GS
	ARMY	
EL	Electronics	AR + MK + EI + GS
GM	General Maintenance	MK + EI + AS + GS
MM	Mechanical Maintenance	NO + AS + MC + EI
OF	Operators & Food	NO + AS + NC + VE
SC	Surveillance & Communication	AR + AS + MC + VE
	MARINE CORPS*	
MM	Mechanical*	AR + EI + MC + AS
	NAVY	
EL	Electronics	AR + MK + EI + GS
ME	Mechanical**	VE + MC + AS
EG	Engineering	MK + AS
MR	Machinery Repair**	AR + MC + AS

* Data were analyzed separately for this Marine Corps composite. (See Appendix A.)
**Data for this composite were included in the overall results, but sample sizes did not permit separate analyses by composite.

Prior Study of the ASVAB Validity Differences by Race and Gender

A limited number of studies have examined gender-related differences in prediction of training and performance outcomes in the military because, historically, relatively few military occupations had enough females to permit meaningful analysis. In the examination of differential gender-related prediction of training success, Booth-Kewley, Foley, and Swanson (1984) found significant differences in slopes for males and females in 2 out of 100 schools (Data Processing and Mess Management, both of which use Verbal [VE] and Arithmetic Reasoning [AR] as the selector composite). In these schools, the slopes were steeper for females; the male regression equation overpredicted final school grades (FSGs) for females in the lower half of the ASVAB 8, 9, and 10 composite score range and underpredicted FSGs for females in the upper half of the score range.

Weltin and Popelka (1983) evaluated the predictive validity of the ASVAB 8, 9, and 10 for Army data using the FSG as the criterion. Female scores were above the male regression line at the lower portion of the composite score range, suggesting possible underprediction for females. The authors did not, however, find significant differences in either the slopes or intercepts to be significant but did find significant differences in the standard errors of estimate for males and females.

Maier and Truss (1984) found the female performance was significantly underpredicted in six Marine Corps training courses. The female underpredictions were especially notable in traditional female occupations, such as administrative clerks and food service handlers. The authors issued a stiff caveat with their findings, however, pointing out the small sample sizes used in their study.

Welsh, Kucinkas, and Curran (1990), in a review of the ASVAB validity data, reported results of two large studies done on Air Force and Navy samples (Wilboum, Valentine, & Ree, 1984; Booth-Kewley, et al. 1984) using the FSG as a criterion in investigations of the predictive equity of the ASVAB 8, 9, and 10 composites. For the Air Force recruit data, the Armed Forces Qualification Test (AFQT) validities for females and males (not corrected for restriction in range) were .42 and .37, respectively. For the Navy, the uncorrected AFQT validities for females and males were .37 and .42. The average AFQT validities for blacks and whites were .20 and .41 in the Air Force samples and .29 and .41 in the Navy samples. The reviewers stated that these differences in mean validities between black and white subgroups from the Wilbourn et al. (1984) study were not consistent with the literature addressing racial differences in prediction for other forms of the ASVAB. They cited studies by Bock and Moore (1984) and information contained in the ASVAB ***Test Manual*** and ***Technical Supplement*** (DoD, 1984a & 1984b). They offered the possible explanation that restriction in range of abilities and consequent reduction in variance of scores of the two subgroups in the Air Force sample could account for reduced correlations for the black subgroup.

McLaughlin, Rossmeissl, Wise, Brandt, and Wang (1984) examined the ASVAB Forms 8, 9, and 10 for ethnicity and gender differences in a large study of Army recruits (N=65,193). The analyses examined the differences between gender and race subgroup specific and common regression lines; the results indicated few or no differences among groups in the regions of the minimum aptitude qualifying scores.

Welsh et al. (1990) concluded that there were mean differences in performance between blacks and whites on the subtests of the ASVAB and that this was consistent with the majority of the literature on tests of mental ability, in particular with the findings of Eitelberg, Laurence, Waters, and Perelman (1984) in the effects of aptitude composites used to select and classify applicants for the American military.

Related Research in the Civilian Sector

Ability tests that are quite similar to the ASVAB have been widely used for selection into civilian occupations, and the issue of their fairness has also been analyzed extensively. In a synthesis on ability testing developed by the National Research Council, Linn (1982) concluded that "there is little evidence for differences in validity coefficients for whites and blacks in civilian employment settings" (p. 373). In a subsequent study of the General Aptitude Test Battery (GATB), Hunter (1983) concluded that apparent race and gender differences in validity were largely or completely due to statistical artifacts. Nonetheless, the issue of the fairness of standardized tests in employment selection persists (Gifford, 1989). Linn and Dunbar (1986) provide a recent summary of differential validity results and references to a wide array of more specific studies.

Methodology for assessing sensitivity and fairness has also received considerable attention in the general literature. Linn and Dunbar (1986) assert that "For purposes of evaluating questions of bias, it is clear that comparisons of correlation coefficients are simply inadequate for the problem" (p. 228). Their primary concern is that correlation coefficients are affected by group heterogeneity and other factors that do not relate to how the selection test is used in predicting an outcome. They conclude that "An adequate evaluation of the question of possible predictive bias demands that regression equations and standard errors of estimate or expectancy tables be compared" (p. 228). Nonetheless, when a National Research Council committee reported its review of the GATB, many of their conclusions about race and gender differences in validity were based on comparison of correlation coefficients (Hartigan & Wigdor, 1989).

The analytic technique known as meta-analysis has contributed significantly to the analysis of test fairness. The literature is characterized by a large number of different studies of the same or related tests used in selection for the same or related jobs. Most studies had sample sizes that were too small or criterion measures that were not sufficiently reliable to detect relatively small differences in predictive relationships. Hunter and Schmidt (1990) provide a summary of meta-analytic methods that have been developed to combine the results of separate studies into a single, more powerful, summary. Their book provides an extensive bibliography for those interested in more detail on the history or variations of this technique.

Approach

A two-phase approach was designed to respond to the request for research to identify more sensitive predictors for technical specialties.[1] The focus of this report is on the first phase: the investigation of the current ASVAB selection composites that involve the technical **subtests** to determine which composites and **subtests** are most in need of improvement with respect to their sensitivity and fairness for all applicant groups and to suggest possible improvements within the context of the current ASVAB.

The basic approach to assessing sensitivity and fairness in the present study was based on analyses of differential prediction. The ***Standards for Educational and Psychological Testing*** (American, ***1985)*** state:

> ***Differential prediction*** is a broad concept that includes the possibility that different prediction equations may be obtained for different demographic groups, for groups that differ in their prior experiences, or for groups that receive different treatments or are involved in different instructional programs.. ..
>
> In a study of differential prediction among groups that differ in their demographics, prior experiences, or treatments, evidence is needed in order to judge whether a particular test use yields different predictions among those groups (e.g., different predictions for males and females). There is differential prediction, and there may be selection bias, if different algorithms (e.g., regression lines) are derived for different groups and if the predictions lead to decisions regarding people from the individual groups that are systematically different from those decisions obtained from the algorithm based on the pooled groups.
>
> The accepted technical definition of predictive bias implies that no bias exists if the predictive relationship of two groups being compared can be adequately described by a common algorithm (e.g., regression line) (p. 12).

The general approach to the assessment of fairness was thus to compare average criterion values for individuals from different groups who had the same score on the selection composite. Sensitivity is a term that is less commonly used in conjunction with selection tests. In the present study, the selection composites were considered sensitive to the extent that differences in composite scores were associated with differences in important criteria. Specifically, sensitivity was operationally defined to be the differences in average criterion scores between individuals who scored one standard deviation above the population mean on the selection composite and individuals who scored at the population mean. As described below, the score range from the population mean to one standard deviation above the mean covered the area of interest in selection for technical specialties. The extent to which the selection composites showed different degrees of sensitivity for males and females and whites and blacks was then examined.

---

[1]A second phase of the investigation of more sensitive measures will involve possible changes to the **ASVAB** battery itself. The Personnel Testing Division of DMDC is currently coordinating a comprehensive review of the contents, administration, and use of the **ASVAB** and is scheduled to submit recommendations for changes to the **ASVAB** in March **1993.** Part of this effort involves examination of possible new subtests: spatial, memory, and psychomotor measures. Evaluation of these new tests will include analyses of their sensitivity and fairness for key applicants from different race and gender groups.

In the evaluation of composites for this report, emphasis was placed on evaluating impact across a broad spectrum of jobs in contrast to the case study approach that was adopted by the GAO. The analyses conducted by the GAO focused on a relatively small number of highly technical Army, Navy, and Air Force specialties. As a consequence, the GAO sample sizes were particularly small when divided into separate sex or ethnic groups. To respond to the GAO, this report takes a somewhat broader perspective and uses relatively large samples for analyses. The objective was to evaluate current selection composites in the context of the entire range of specialties for which they are used and to maximize the statistical power to detect differences by combining results across jobs where appropriate. Except for this broader focus, the criterion measures and samples used in the present study closely paralleled those reported by the GAO.

Data

Three different data sets were used in the analyses reported here. Navy and Air Force data on training success and Army data on Skill Qualification Test (SQT) results were analyzed. For the first two data sets, training courses were the primary unit of analysis, and course grades were the measure of success in training that was analyzed. For the SQT data, each distinct form of the SQT (generally one per year per specialty) was analyzed separately, and the score on that form was used as a measure of success on the job.

Navy Training Data

Data were collected from Navy training courses in Type A schools over the period 1989 to 1990. For the Navy courses included in this study, Final School Grade (FSG) was the criterion measure. In Navy training data, FSG generally represents an arithmetic average or a weighted sum of grades earned on daily and/or weekly quizzes, measures of hands-on performance and practical proficiency, and the score on a final comprehensive exam.

Data on performance in technical schools were included in the present analyses. In this case, technical schools were defined as those for which one or more of the ASVAB technical subtests was included in the selection composites. The three subtests classified as technical are Auto and Shop Information (AS), Electronics Information (EI), and Mechanical Comprehension (MC). All courses with at least 40 blacks and at least 40 whites were used in the analyses of race differences. Similarly, all courses with at least 40 females and at least 40 males were used in the analyses of sex differences. Appendix B on page 32 lists the Navy specialties and sample sizes included in the present analyses.

Air Force Training Data

Data were collected from Air Force technical training schools and courses from approximately January 1985 until June 1988. For this study, technical schools were defined as those whose selection composite included one or more of the ASVAB technical subtests (AS, EI, or MC). All courses for which at least 40 blacks and 40 whites had valid data were used in the analysis of race differences, and all courses for which at least 40 males and 40 females had valid data were used in the analyses of sex differences.

The criterion measure was the FSG. This measure, like the Navy FSG, often represents an aggregation of multiple-choice tests. The Air Force employs performance checks during training that are analogous to hands-on tests used in Navy training schools. In normal practice, Air Force trainees may take the performance checks several times. There is no information in these data sets on how many times a given trainee has taken the performance check (Ree & Earles, 1990). FSGs for the Air Force range from approximately 60 (lowest) to 99 (highest). Appendix C, beginning on page 33, lists the Air Force specialties included in the present analyses.

Army SQT Data

From 1978 until it was canceled in 1990, the SQT program in the Army was the most extensive job-proficiency testing program in history. As originally implemented in 1978, SQTs were designed to be criterion-referenced tests of job proficiency. Each SQT had three components: written component, hands-on component, and performance certification component (when a soldier's supervisor would observe the soldier performing a certain task during normal working hours and score the soldier as successful or unsuccessful at performing the task). In addition, SQTs were originally designed to measure both the individual soldier's job proficiency and the training effectiveness (Maier & Hirshfeld, 1978).

There are more than 250 Military Occupational Specialties (MOS) in the Army, each of which has soldiers in one to five skill levels. Skill level 1 refers to soldiers in pay grades E-1 through E-4; skill level 2 soldiers are in pay grade E-5; skill level 3 soldiers are in pay grade E-6; skill level 4 soldiers are in pay grade E-7; and skill level 5 soldiers are in pay grades E-8 and E-9. Soldiers were required to take the SQT annually in their MOS and skill level until they received a GO (passing 80% of the tasks tested on the SQT) on the test.

In 1983 the SQT program underwent a major revision resulting in the Individual Training and Evaluation Program. The training effectiveness evaluation, hands-on testing, and performance certification were separated from the job proficiency portion of the SQT program. Local commanders selected tasks for evaluation that supported their unit's mission and used the results to guide training needs. The Common Task Test (CTT) was developed by the Training and Doctrine Command (TRADOC) and was administered to soldiers in skill levels one through four in all MOS once a year. The CTT was composed of tasks tested primarily in the hands-on mode. Results of the **CTT** were provided to TRADOC and to local commanders to be used as a factor in determining training needs.

After 1983, the SQT became a task-based written test designed to measure job proficiency of individual soldiers. Soldiers with 11 months or more of service were required to take the SQT annually if the test was available in their MOS and skill level. Compilation of the 1988-1989 SQT records show that more than 90% of the skill level 1 MOS had the SQT in at least one of those two years, and about 90% of skill level 1 soldiers took one or more SQTs during that period. Results from skill level 1 and skill level 2 SQTs were used in making promotion decisions for pay grades E5 and E6 respectively.

Specific guidance for developing the SQT was provided to test developers (TRADOC Regulation 351-2). This guidance was in accordance with standard test development procedures and includes the minimum and maximum number of tasks to be tested, the use of random and random-strat ed selection of tasks, tryout procedures, security, etc. Tasks eligible to be tested are contained in the Soldier's Manual appropriate to each MOS and skill level.

The samples used in the current analyses are part of a large ASVAB validity study currently underway in the Army. The current samples were limited to the task-based written test, skill level 1 SQT. The sample was further limited to soldiers who had originally taken the ASVAB in its current format (ASVAB forms 8-17). Entry ASVAB scores for 1981-1988 accessions were matched against the SQT records for 1985-1990. All SQT/year samples containing at least 50 soldiers were retained, resulting in 1,004 analysis samples in 204 of the potential 242 entry level MOS.

In the current analyses, all samples with at least 40 blacks and 40 whites were used in the analyses of race differences. Similarly, all samples with at least 40 females and 40 males were used in the analyses of sex differences. The samples were further restricted to the MOS for which the ASVAB selection composite included one of the technical subtests (EI or AS). Appendix D, beginning on page 35, lists the Army specialties and sample sizes included in these analyses.

Marine Corps Hands-On Performance Data

Data on Marine Corps mechanical specialties collected by the Job Performance Project were analyzed separately by researchers from the Center for Naval Analyses. The criterion measure used was the percentage of steps performed correctly in a representative sample of job tasks. The high fidelity nature of the criterion used made these analyses particularly important, but the samples used in these analyses were too small to allow a meaningful contribution to pooled analyses. Consequently, results from analyses of these data are reported separately in Appendix A, beginning on page 29.

The ASVAB Scores

The ASVAB scores of record were analyzed for each of the samples described above. As indicated, the samples were restricted to specialties for which technical **subtests** were used in selection. Table 3 below shows means, standard deviations, reliability estimates (coefficient alpha), and standard errors of measurement for the three technical subtests.

The data shown are from a recent administration of the Reference Form (Form 8a) to a sample of new recruits during a preliminary calibration of new forms (Forms 20, 21, and 22). Recruits were used in this example rather than applicants so that the variation in abilities would be more comparable across race and gender groups, and thus, reliabilities could be more meaningfully compared. Reliabilities were not corrected for restriction in range and so are considerably less than standard estimates of reliability for the youth population as a whole.

As shown in Table 3, there were smaller reliability estimates for females and blacks in comparison to the total sample. Nearly all of the difference is due to differences in standard deviations, so the standard errors of measurement are quite similar. Differences in standard errors were due, in part, to the fact that females and blacks more frequently scored at the lower end of the scale where error of measurement tends to be greater due to a greater frequency of guessing.

Table 3
Descriptive Statistics, Reliabilities, and Errors of Meaurement for the Technical Subtest Number Correct Scores

Statistic	Subgroup	n	AS	Subtest MC	EI
Mean	Total	2418	15.5	15.6	12.3
	Female	293	11.6	12.1	9.7
	Black	378	11.5	11.8	10.3
	Hisp.	165	13.9	14.6	10.9
S.D.	Total	2418	4.5	4.7	3.4
	Female	293	3.5	4.0	3.0
	Black	378	3.4	3.9	3.2
	Hisp.	165	3.8	4.4	3.5
REL.	Total	2418	0.77	0.79	0.69
	Female	293	0.59	0.68	0.58
	Black	378	0.57	0.67	0.61
	Hisp.	165	0.66	0.75	0.69
SEM	Total	2418	2.2	2.2	1.9
	Female	293	2.3	2.3	2.0
	Black	378	2.3	2.3	2.0
	Hisp.	165	2.2	2.2	1.9

Analyses

The data analyses were conducted in three stages. The first stage consisted of data edits and adjustments. In the second stage, separate analyses were performed for each distinct sample. In the final stage, the results were aggregated across samples yielding summary results for each of the ASVAB composites analyzed and also for all of these composites combined. Appendix E, beginning on page 41, provides details, formulas, and examples for each step in the analyses.

Data Edits and Adjustments

For the most part, the data files were already clean and complete. A small number of cases missing either predictor or criterion data were deleted. The one edit of substance eliminated all cases where the ASVAB composite score of record was below the current selection cutoff for the specialty. The majority of these cases had been granted waivers and allowed to enter their specialty with ASVAB scores that would not otherwise have qualified. These individuals were likely to possess other unmeasured qualities that led to a waiver; therefore, they were not strictly comparable to individuals who came in normally. It was also possible that their ASVAB scores were in error, which would also support exclusion from the present analyses. In all, about 5% of the initial records were eliminated for this reason.

For samples with training criteria, some data were available on individuals who did not successfully complete their training. The prediction of training completion is more important than the prediction of differences in final grades among those who do complete. For this reason, information on training failures was retained wherever possible. In most cases, no appropriate FSG was available for these cases, so a final grade was imputed. The procedure used assumed that the overall distribution of final grades (for both successes and failures) was approximately normal with successes scoring above a cut score and failures scoring below the cut score. The proportion passing the course was used to estimate where the cut score would be on the normal curve that was fit to the observed mean and standard deviation of scores for those who passed. The mean score for those below the cut point was computed and assigned to all of the failures.

In addition to screening out inappropriate cases and imputing scores for training failures, adjustments to the criterion scores were computed to improve comparability across specialties. The nature of the criterion measure differed somewhat (primarily in terms of level or difficulty) across specialties within each Service and differed more considerably across the Services. In general, it took a higher level of ability to receive a given score in a very selective specialty than it did in a less selective specialty. For the basic comparisons to be made, the scaling of the criterion variable within each sample was irrelevant. As described below, analyses were performed separately for each specialty sample. The statistics that were computed and aggregated across samples were t statistics that would be unchanged by any linear transformation of the criterion scale. Nonetheless, a linear transformation of the criterion scales for each sample was performed to reduce differences due to sample selectivity and related criterion difficulty. The goal in making these transformations was to minimize the possibility that graphs of prediction curves for each group separately might be distorted by complex interactions between the scaling, the curvature, and perhaps other factors associated with the prediction functions for each separate sample. Differences due to variation in the reliability or other aspects of the criterion could not be eliminated, as insufficient information was available on the distinct psychometric properties of each measure.

The criterion scores were adjusted so that if the criterion for each training course or SQT were available for the entire youth population, the (expected) means and standard deviations for each criterion would be the same. The adjustment made was the reverse of the adjustment that is typically made to correct for restriction in range due to selection. In the normal case, job specific sample means and correlations are adjusted to estimate the corresponding statistics in the youth population as a whole using the multivariate range restriction procedure developed by Lawley (see Lord & Novick, 1968, p. 147). In the present case, the criterion scales were adjusted so that the estimated youth population mean and standard deviation would be the same for each sample. A mean of 85 with a standard deviation of 5 was initially used with the Navy and Air Force training data, and a mean of 70 with a standard deviation of 10 was initially used with the Army SQT data. These were close to the observed values and minimized the adjustments that were made. Subsequently, both the predictor and criterion variables were restandardized to have a mean of zero and a standard deviation of one in the youth population.

The specific procedure used for each sample was to develop a regression equation for predicting the criterion from the ASVAB subtest scores, estimate a youth population mean on the original scale by substituting population means of 50 for each ASVAB subtest for the sample subtest means, estimate the youth population variance on the original criterion measure using the multivariate correction referenced above, and develop a linear transformation of the criterion scale values that transformed the estimated youth population means and standard deviations to the target values.

Individual Sample Analyses

Analyses of the individual samples were designed to address two key questions. The first question concerned the sensitivity of the selection composite used with the specialty in question. The initial concern expressed in the GAO report was with the most selective specialties and, for this reason, focus was concentrated on the upper end of the selection test scale. The operational definition used for ***sensitivity*** was the ***difference in expected training or job success between an individual who scored at the youth population mean and an individual who scored one standard deviation above the youth population mean.*** Note that this definition is equivalent to the slope of the regression line in a linear regression with standardized predictor scores. The selection composite is thus a sensitive predictor if differences in test scores are associated with important differences in job outcomes.

As an alternate indicator of sensitivity, the prediction error was examined to see if the selection composite provided a more accurate prediction for some applicant groups than for others. When the standard error of prediction was small, then the selection composite was also considered to be an accurate predictor of the outcome in question.

Correlations were considered an inappropriate measure of sensitivity, even when adjusted for differences due to restriction of range, because correlations depend heavily on the heterogeneity of the sample with respect to both predictor and criterion measures, and adjustments for differences in heterogeneity may undercorrect in many cases. In addition, the relationship of the predictor and criterion measures may not be linear, as was found in the present analyses.

The second question addressed in the analyses concerned fairness. The operational defmition used for ***fairness*** was the ***extent to which individuals at a given test score level had the same expected performance level regardless of race or gender,*** following the generally accepted definition of fairness (Cleary, 1968). When the test score level and expected performance level were even regardless of race or gender, then the test was judged fair for all groups.

In addressing both questions, a model of the relationship of the criterion measures to the predictor (selection test) was required. There were too few individuals in each applicant group who scored exactly at the youth population mean or exactly one standard deviation above it to estimate sensitivity reliably. Similarly, there were too few examinees at any given score level to analyze each score level separately with respect to fairness. Consequently, some model of the relationship between predictor levels and expected outcomes was needed.

It is common to adopt a linear model of the relationship of the criterion measure to the selection test and to perform linear regression in assessing this relationship. A linear model has a constant slope implying that the prediction is equally sensitive across all score levels. By contrast, a quadratic or higher order polynomial model would allow for differences in slope or sensitivity at different predictor score levels.

Since sensitivity was a key issue in these analyses, a test for nonlinear effects was run before deciding whether to adopt a linear model. The data was pooled by selection composite. With a separate test for each individual sample, limited sample sizes might preclude an accurate answer in many cases and result in hundreds of tests with some significant results due to chance factors. Further, with all data pooled into a single analysis, true differences in the nature of the relationships for different selection composites, and also for the different types of criterion measures (training versus on-the-job), might have been masked.

As described in the Results section in this report, a quadratic regression model was adopted. In analyzing fairness, differences in predicted criterion scores over the selection test range from one standard deviation below the youth population mean to 'one standard deviation above the population mean were looked at. (Virtually all selection decisions are made in this range.)

One other issue in the analyses was the effect of the restriction in range on the results. Outcome data were only available on individuals who had passed all selection screens and been enlisted into the military. In addition, the Army SQT data were only available on individuals who had successfully completed training and remained on the job for a period of time. The objective was, however, to generalize the fmdings from the specific samples analyzed to the population of applicants. The samples studied had significantly less variation in the ASVAB scores compared to all applicants or to the 1980 youth population, and correlations would be significantly attenuated by this difference. Explicit selection on the predictor being analyzed would not affect regression lines so long as additional selection factors were not correlated with both the predictor and the criterion. Unfortunately, it was not possible to develop detailed models of implicit selection factors. To the extent that they existed, it seems likely that the implicit selection factors would have had a positive relationship with both the predictor and criterion. (Individuals with high predictor scores **and/or** high criterion scores would be more likely to remain in the sample.) In this case, the uncorrected results would understate the significance of the relationship between predictor and criterion measures, overall and for each race and gender group. In this sense, the unadjusted values are conservative in that they are likely to be a lower bound.

Methods for Aggregating Results

The analyses of sensitivity and fairness in each of the individual samples led to hundreds of answers to the question of race and sex differences. It was necessary to develop an overall assessment of each different selection composite and of the technical portion of the ASVAB as a whole. The general approach was to compute estimates of key subgroup differences in each sample and then to compute weighted averages of these differences across samples and test whether the weighted averages of the differences were significantly different from zero. This approach both summarized the results from hundreds of separate samples and allowed for a much more powerful test of differences, owing to the very large number of observations in the combined samples.

The significance tests used with the overall results were based on a normal approximation. Given the large number of samples that were combined (more than 100 for the gender analyses and more than 300 for the race analyses), the central limit theorem ensured that the mean of the individual t statistics would have a nearly **normal** distribution. In addition, while the exact degrees of freedom for the aggregate statistic was not computed, it was very large (hundreds, if not thousands), so treating the aggregate statistic divided by its standard error as a z statistic was entirely appropriate. Appendix E provides details and examples on the aggregation procedures.

The specific statistics analyzed to test for differences related to gender or race were

- ***sensitivity:*** the predicted criterion score at one standard deviation above the youth population mean on the predictor minus the predicted criterion score at the youth population mean (for linear models, this would be equivalent to the difference in slopes);

- ***error of prediction:*** the root mean square error from the (quadratic) regression analysis; and

- ***predicted criterion scores:*** at five key points on the predictor scale (ranging from one standard deviation below the youth population mean to one standard deviation above the youth population mean), used in assessing fairness.

Several different procedures for pooling results across samples were used. The initial approach was to weight each difference by the inverse of the standard error of the statistic. In this way, difference estimates from small samples that were not very accurate (had large standard errors) would not get very much weight (the inverse of the standard error) in comparison to statistics from samples that provided more accurate estimates. This approach was equivalent to taking a simple average of t-values (differences divided by their standard errors) across the samples. Since t-values are independent of the measurement scale, this approach had the advantage of eliminating the issue of the equivalence of the criterion scales across samples.

Hedges and Olkin (1985) show that the most accurate estimate of a statistic across multiple samples is obtained when the individual sample statistics are weighted by the inverse of the square of the standard error of the statistic rather than by the inverse of the standard error. Results using such optimal weights also were examined. The composite standard errors for testing for mean group differences were slightly smaller, but the effect size estimates were quite similar, and there were no differences in conclusions.

For a given sample, each of the statistics of interest had a different weight under both the t-value and optimal weighting schemes. Differences at the lower end of the predictor scale would have smaller standard errors and larger weights for samples that included more lower-scoring incumbents in comparison to equal size samples with higher-scoring incumbents. The aggregate test for differences at the low end of the predictor scale gave more weight to lower scoring samples, and the test for differences at the high end of the predictor scale gave more weight to higher scoring samples. For purposes of assessing differences at each different predictor level, this differential weighting was entirely appropriate. When it came time to plot the complete regression curves for each group, the use of different sample weights for different predictor levels might have led to significant interaction effects. Another set of weighted averages was computed by using the inverse of the standard error of criterion differences at the youth population mean as the weight (population mean difference weights) for all of the statistics analyzed. Again, this led to very similar estimates of effect sizes and no differences in conclusions. Finally, unweighted averages also were computed for comparison purposes.

In this report, the original t-value weights are reported for the individual statistics, and the population mean difference weights were used in preparing the graphical displays of the regression curves. In the graphical displays, linear interpolation was used to fill in the curves between the criterion levels estimated for the five key predictor levels.

For each sample, the criterion level for each predictor level was estimated as a linear composite of the three regression parameter estimates (intercept, linear, and quadratic coefficient). As described in Appendix E, a standard error for each predicted value was estimated using estimates of the variances and covariances of the parameter estimates. Standard errors for the aggregate values were estimated using a weighted combination of the squares of the standard errors for the individual sample values. Variability in the estimates of the weights for each sample was not considered in estimating confidence bounds. The approach was appropriate for a model in which the weights are held fixed at their current value and not re-estimated in each replicate sample. Estimation of confidence bounds for a model in which the weights were also re-estimated in each replicate sample would have been quite complex and, since the weighting of the individual samples was not the question of interest, was judged unnecessary. The confidence bounds also do not include variability associated with the criterion scale adjustments. If separate criterion scale adjustments were estimated for each replication, the variability across replications, and hence the confidence bounds, would be somewhat greater.

Since the criterion scaling was largely irrevelant to the issues at hand, estimating confidence bounds for the condition that the scaling was held constant across replications was judged to be most appropriate.

In addition to an overall aggregation of results, separate aggregations were computed for each different selection composite for which data on at least 400 members of each applicant group were available. A cutoff of 400 was selected as this leads to confidence bounds for mean estimates of .1 standard deviation or less, a level of accuracy judged adequate to support conclusions about the predictor-criterion relationships. Aggregate results were not analyzed for two of the composites originally identified for inclusion in the study due to insufficient sample size. The small amount of data available on specialties using these composites was, however, included in the overall aggregate results.

Results

Tests for Linearity

Table 4 (a and b), on page 16, shows the results of the analyses used to test for the linearity of the relationship between the predictor and criterion variables. Linear through quartic predictor terms and subgroup main effects and interactions were included in the analyses. In these analyses, data were pooled across all of the samples that had the same selection (predictor) composite. Table 4 shows the F statistic testing the significance for each term controlling for the effects of all preceding terms, but not for the effects of the terms that follow. The individual F statistics have one degree of freedom in the numerator and a large number (> 100) of degrees of freedom in the denominator. The critical value for an alpha of .05 for such statistics is about 5.1. Since the F statistic is a ratio, harmonic means (across composites) were used as an indicator of the average effect of each term. The results indicate the clear statistical significance of linear and quadratic terms and of subgroup main effects for the majority of the composites analyzed. Some of the remaining terms were significant for some of the composite samples, but the overall means were quite close to one, the value expected under the null hypothesis (no effect). The significance of the higher order terms in some samples may have resulted, in part, from complex interactions between samples and predictor score distributions that would not have held up when separate analyses were performed for each sample. Based on the results shown in Table 4, it was decided to proceed with quadratic regressions even though, as indicated by the relative F values, the practical significance of the quadratic term was quite small. The relative cost of over-specifying the prediction model was minimal: a few extra degrees of freedom (two per sample) resulted in an essentially straight line. The cost of under-specifying the prediction model might have been much greater.

Table 4a
Polynomial Regression by Race: F Values for Successive Terms

Composite	P	P2	S	P3	P4	SxP	SxP2	SxP3
AF-E	2547.24	0.03	18.20	13.80	0.01	3.24	0.86	1.43
AF-M	1674.28	26.01	9.84	3.78	5.02	0.30	3.37	0.05
AR-EL	12357.17	14.01	270.44	23.84	2.24	1.34	0.08	0.19
AR-GM	10053.99	57.55	121.57	6.15	3.24	2.21	0.77	3.30
AR-MM	30907.80	429.39	738.78	0.00	6.32	1.38	6.27	3.25
AR-OF	16590.25	4.37	466.64	0.16	0.99	25.05	0.11	1.53
AR-SC	3951.43	41.34	21.98	0.19	0.09	1.53	6.94	0.01
NA-EL	1859.13	23.95	13.92	5.67	14.76	0.85	0.12	0.00
NA-EG	1484.28	21.33	15.56	4.03	2.30	1.84	1.42	9.99
NA-ME	160.89	1.25	2.94	0.01	0.26	0.02	0.04	2.85
NA-MR	6799.29	147.06	58.13	5.05	0.85	6.95	1.90	1.33
Hrm Mean	3907.30	14.16	45.31	1.01	1.04	1.37	0.66	0.48

Table 4b
Polynomial Regression by Sex: F Values for Successive Terms

Composite	P	P2	S	P3	P4	SxP	SxP2	SxP3
AF-E	3762.85	1.58	16.13	12.62	1.18	0.09	0.64	0.18
AF-M	1401.54	15.71	1.43	3.18	6.06	7.66	0.09	0.09
AR-EL	12471.62	3.60	139.32	31.73	1.17	62.62	2.60	2.99
AR-GM	5425.80	86.81	65.81	11.74	2.44	1.45	1.49	0.01
AR-MM	21113.97	335.47	913.90	3.25	0.01	21.31	2.85	0.52
AR-OF	13081.86	4.46	0.80	5.73	3.89	9.58	4.44	0.83
AR-SC	4073.04	44.39	36.06	0.15	0.16	0.64	2.28	0.11
NA-EL	4747.16	34.26	3.75	2.13	15.63	0.17	'1.50	0.74
NA-EG	1377.38	18.86	11.33	4.17	2.35	0.05	3.23	0.14
NA-ME	38.28	0.21	3.82	1.55	0.15	0.23	1.48	0.19
NA-MR	359.35	3.25	0.06	0.74	2.43	1.86	0.11	0.00
Hrm Mean	2596.16	10.77	9.46	3.25	1.05	1.36	1.16	0.16

P, P2, P3, and P4 are the linear, quadratic, cubic, and quartic terms for the predictor and S denotes subgroup effects. Each element in the table is an F statistic with one degree of freedom at a large number (> 100) of degrees of freedom in the denominator. The critical value for such an F statistic is about **5.1** (alpna = **.05).**

Aggregation of Results

Table 5 (a and b) below shows the overall means and standard deviations across samples of the t-values used to summarize the differences of interest. As described in Appendix E, an approximation that does not assume equal underlying variances was used; consequently, the degrees of freedom depend on the ratio of the underlying variances as well as the sample sizes. In all cases, the degrees of freedom were greater than the smaller of the two samples minus one, and so at least 39. Even at this minimum degrees of freedom, the variance of the t statistic is not more than 10 percent greater than one, and so, under the null hypothesis of no differences by race or gender, the t-values would have a mean of zero and a standard deviation of close to one. The significance of the mean differences is discussed below. It is interesting to note that the standard deviations were only slightly larger than one. Systematic variability across samples in the size of mean differences would increase the overall variation in the t-values above one. The finding that the variance of the t-values was only slightly above one suggests that such systematic differences were small.

Table 5a
Distribution of T-Values Across Samples by Race*

Statistic	Mean	Standard Deviation	Minimum	Maximum
Sensitivity	-0.212	1.066	-4.134	2.944
Perf. at -1.0 sd	-0.156	1.074	-2.646	3.024
Perf. at -0.5 sd	-0.449	1.288	-4.581	3.058
Perf. at the mean	-0.899	1.761	-6.525	4.251
Perf. at +0.5 sd	-1.073	1.555	-7.885	3.271
Perf. at +1.0 sd	-0.775	1.184	-4.759	3.025
Prediction Error	0.046	1.432	-8.003	4.829

***Results by Race (338 Samples)**

Table 5b
Distribution of T-Values Across Samples by Sex**

Statistic	Mean	Standard Deviation	Minimum	Maximum
Sensitivity	0.343	1.008	-2.024	2.965
Perf. at -1.0 sd	-0.205	1.204	-5.092	2.575
Perf. at -0.5 sd	-0.564	2.140	-10.941	5.212
Perf. at the mean	-0.650	2.775	-11.843	5.544
Perf. at +0.5 sd	-0.164	2.139	-8.535	5.099
Perf. at +1.0 sd	0.090	1.516	-5.123	3.708
Prediction Error	-0.306	1.367	-5.436	2.957

****Results by Sex (166 Samples)**
Difference (focal - reference group values)

Differences in Sensitivity

Table 6 (a and b) on page 19 shows the estimates of sensitivity differences by race and sex respectively. In these and subsequent analyses, both selection test and criterion scores were standardized to have a mean of zero and a standard deviation of one in the youth population. In this metric, the sensitivity measure is analogous to an estimate of the correlation of predictor and criterion scores in the youth population as a whole. (The sensitivity measure would be identical to the correlation, corrected for restriction in range, if a linear model were used.)

The sensitivity measures are quite high for all groups. Overall, each group shows over a half standard deviation gain in the criterion measure for a one standard deviation increment in selection composite level. In the aggregate, the selection composites are quite sensitive in identifying potentially able performers. The results by sex are quite different from the results by race. Here, the ASVAB technical composites were found to be more sensitive predictors for females than for males. This result was also found for most of the individual composites, although the differences were significant for only about half of the composites.

In the aggregate, the sensitivity measures were greater for whites than for blacks, although the differences are only statistically significant in relatively large samples. The Navy's EL composite was the one composite that showed greater sensitivity for blacks than for whites, although this difference was not statistically significant.

Standard Error of Prediction

Differences between blacks and whites in terms of standard error of prediction were mixed. (See Table 7a on page 20.) For two composites there was a slight but statistically significant difference with smaller prediction errors for whites. For two other composites the opposite was true. Overall, there was not a significant difference.

The sex differences in prediction errors were quite consistent with the sensitivity differences. (See Table 7b on page 20.) Overall, prediction errors were significantly smaller in the female samples. Small but significant differences in the same direction were found for three of the individual composites. There were no composites for which the prediction errors were significantly smaller for males.

Table 6a
Sensitivity Measures by Race

Composite	No. of Samples	Total Cases Blacks	Total Cases Whites	Sensitivity Blacks	Sensitivity Whites	Sensitivity Diff.	t
Total	338	95,080	281,063	0.56	0.62	-0.06	-3.9**
Air Force							
E	17	1,121	11,070	0.67	0.75	-0.08	-0.6
M	11	853	8,682	0.40	0.47	-0.07	-0.7
Army							
EL	70	19,460	39,905	0.50	0.56	-0.06	-2.3*
GM	78	11,759	34,876	0.56	0.61	-0.05	-1.3
MM	75	17,847	71,485	0.73	0.77	-0.04	-1.1
OF	53	33,684	63,429	0.53	0.62	-0.10	-2.6**
SC	14	6,636	19,930	0.56	0.60	-0.04	-0.9
Navy							
EL	11	1,754	18,087	0.52	0.46	0.06	0.5
EG	4	1,544	11,096	0.41	0.48	-0.07	-0.9

Table 6b
Sensitivity Measures by Sex

Composite	No. of Samples	Total Cases Females	Total Cases Males	Sensitivity Females	Sensitivity Males	Sensitivity Diff.	t
Total	166	33,017	249,712	0.71	0.61	0.09	4.3**
Air Force							
E	17	1,580	10,113	0.56	0.72	-0.16	-1.4
M	8	750	7,742	0.67	0.43	0.24	2.2*
Army							
EL	43	6,981	45,023	0.65	0.57	0.08	2.2*
GM	26	2,008	23,032	0.70	0.60	0.11	1.5
MM	23	3,738	49,241	0.77	0.73	0.04	0.6
OF	22	11,233	65,286	0.86	0.70	0.16	3.7**
SC	15	4,455	23,484	0.74	0.58	0.16	3.2**
Navy							
EL	6	1,607	12,405	0.56	0.53	0.04	0.3
EG	3	466	12,186	0.74	0.58	0.16	0.9

* - difference significant at the .05 (two-tail) level
** - difference significant at the .01 (two-tail) level

Table 7a
Standard Error of Prediction by Race

Composite	No. of Samples	Total Cases Blacks	Total Cases Whites	Standard Error of Prediction Black	Standard Error of Prediction White	Diff.	t
Total	338	95,080	281,063	0.78	0.77	0.01	1.1
Air Force							
E	17	1,121	11,070	0.56	0.55	0.02	0.8
M	11	853	8,682	0.65	0.68	-0.03	-1.1
Army							
EL	70	19,460	39,905	0.78	0.80	-0.02	-2.7**
GM	78	11,759	34,876	0.78	0.80	-0.01	-1.4
MM	75	17,847	71,485	0.78	0.74	0.04	5.0**
OF	53	33,684	63,429	0.86	0.84	0.02	2.6**
SC	14	6,636	19,930	0.85	0.85	-0.00	-0.0
Navy							
EL	11	1,754	18,087	0.62	0.63	-0.01	-0.8
EG	4	1,544	11,096	0.75	0.79	-0.04	-2.0*

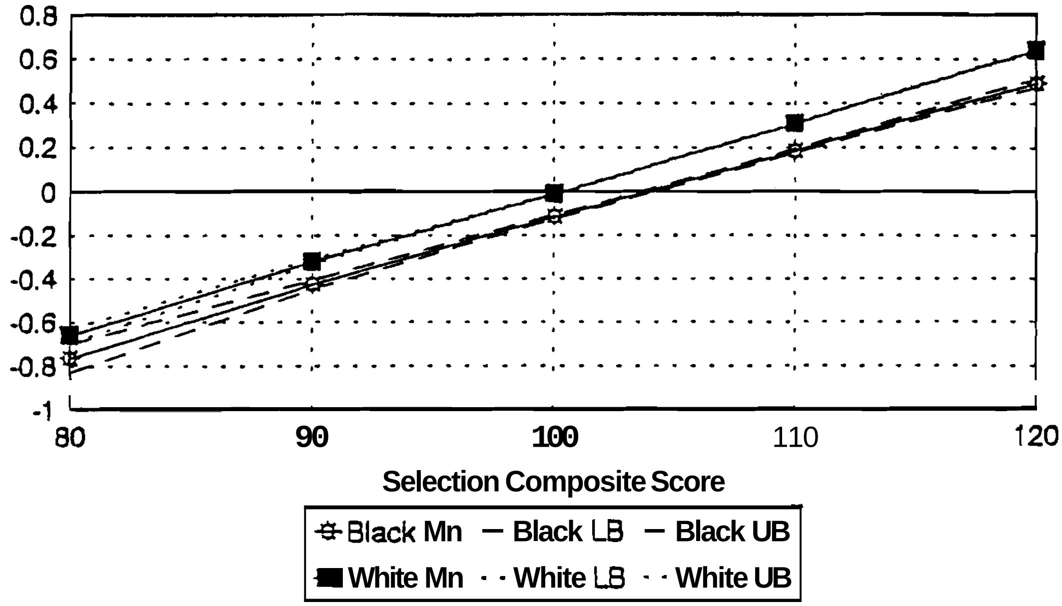
Table 7b
Standard Error of Prediction by Sex

Composite	No. of Samples	Total Cases Females	Total Cases Males	Standard Error of Prediction Females	Standard Error of Prediction Males	Diff.	t
Total	166	33,017	249,712	0.75	0.78	-0.02	-3.9**
Air Force							
E	17	1,580	10,113	0.53	0.53	-0.00	-0.2
M	8	750	7,742	0.69	0.69	0.00	0.0
Army							
EL	43	6,981	45,023	0.78	0.81	-0.03	-2.9**
GM	26	2,008	23,032	0.77	0.82	-0.05	-2.7**
MM	23	3,738	49,241	0.74	0.77	-0.03	-2.1*
OF	22	11,233	65,286	0.83	0.84	-0.01	-0.8
SC	15	4,455	23,484	0.85	0.85	-0.01	-0.5
Navy							
EL	6	1,607	12,405	0.63	0.64	-0.01	-0.6
EG	3	466	12,186	0.77	0.77	0.00	0.1

* - difference significant at the .05 (two-tail) level
** - difference significant at the .01 (two tail) level

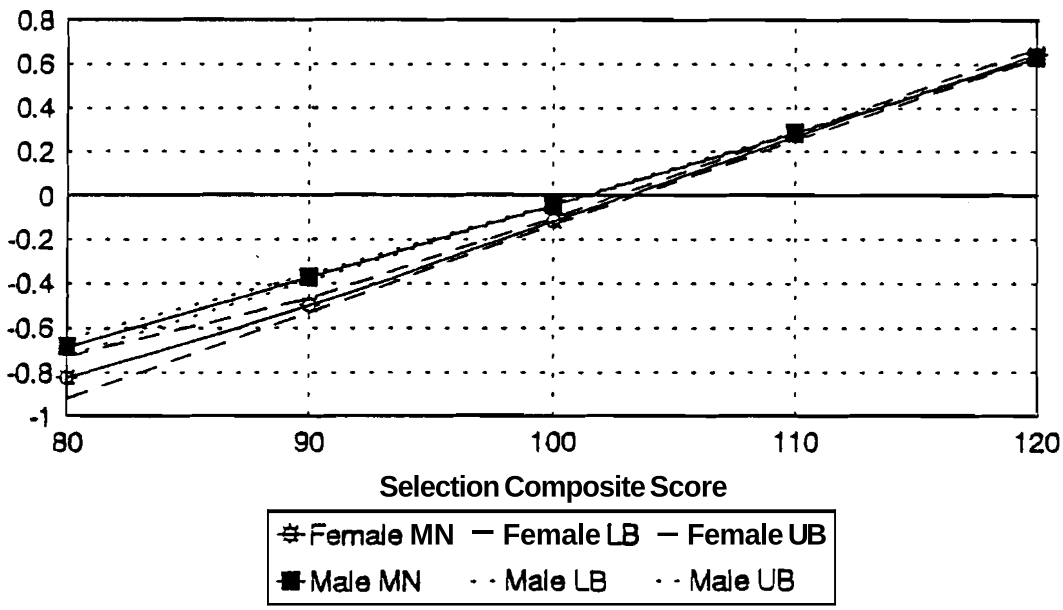
Fairness

Figures 1 and 2 below show predicted criterion levels at key selection composite levels by race and sex for all samples combined.



Based on 338 Samples with a Total of 95,080 Blacks and 281,063 Whites

Figure 1. Predicted Performance by Race: Pooled Results for All Composites



Based on 167 Samples with a Total of 33,104 Females and 249,980 Males

Figure 2. Predicted Performance by Sex: Pooled Results for All Composites

Table 8 (a and b), below and on page 23, shows the statistical comparison of differences in these predicted criterion levels.

Table 8a
Prediction Differences at Key Points by Race

	Prediction at -1.0 8.d.				Prediction at -0.5 8.d.				Prediction at Pop. Mean			
Comp.	Black	White	Diff	t	Black	White	Diff	t	Black	White	Diff	t
Total	-0.76	-0.65	-0.11	-2.8**	-0.43	-0.32	-0.11	-8.2**	-0.12	-0.02	-0.10	-16.5**
Air Force												
E	-0.77	-0.54	-0.23	-0.4	-0.34	-0.25	-0.10	-0.3	0.05	0.09	-0.04	-0.3
M	-1.14	-0.11	-1.03	-2.2"	-0.50	-0.04	-0.46	-2.2"	-0.01	0.10	-0.11	-1.9
Army												
EL	-0.82	-0.57	-0.25	-3.3**	-0.41	-0.25	-0.16	-5.5**	-0.07	0.05	-0.12	-9.7**
GM	-0.60	-0.53	-0.07	-1.0	-0.35	-0.25	-0.10	-4.4**	-0.10	0.01	-0.11	-7.4**
MM	-1.00	-1.01	0.02	0.2	-0.59	-0.47	-0.12	-3.9**	-0.18	-0.03	-0.16	-10.4**
OF	-0.77	-0.75	-0.01	-0.2	-0.47	-0.40	-0.07	-3.0**	-0.20	-0.08	-0.13	-9.9**
SC	-0.86	-0.58	-0.28	-2.5*	-0.38	-0.34	-0.03	-0.8	-0.01	-0.09	0.08	3.8**
Navy												
EL	0.32	0.21	0.11	0.2	0.08	0.08	0.01	0.0	0.08	0.13	-0.05	-0.7
EG	0.31	-0.42	0.73	1.5	-0.11	-0.29	0.17	1.0	-0.23	-0.10	-0.13	-2.8**

	Prediction at +0.5 s.d.				Prediction at +1.0 s.d.			
Comp.	Black	White	Diff	t	Black	White	Diff	t
Total	0.18	0.31	-0.13	-19.8**	0.49	0.64	-0.15	-14.27**
Air Force								
E	0.43	0.49	-0.06	-1.7	0.74	0.86	-0.12	-4.1**
M	0.27	0.31	-0.04	-1.3	0.43	0.56	-0.13	-2.6**
Army								
EL	0.22	0.33	-0.12	-8.6**	0.45	0.61	-0.16	-8.3**
GM	0.17	0.30	-0.13	-6.8**	0.47	0.62	-0.16	-4.6**
MM	0.15	0.36	-0.21	-16.9**	0.53	0.71	-0.18	-7.9**
OF	0.06	0.22	-0.16	-11.8**	0.33	0.53	-0.20	-6.4**
SC	0.29	0.19	0.10	5.0**	0.52	0.51	0.01	0.4
Navy								
EL	0.32	0.42	-0.09	-2.5*	0.72	0.75	-0.03	-1.0
EG	-0.03	0.15	-0.19	-2.8**	0.50	0.48	0.02	0.2

* - difference significant at the .05 (two-tail) level;
** - difference significant at the .01 (two-tail) level

Table 8b
Prediction Differences at Key Points by Sex

Comp.	Prediction at -1.0 s.d. Fem.	Male	Diff	t	Prediction at -0.5 s.d. Fem.	Male	Diff	t	Prediction at Pop. Mean Fem.	Male	Diff	t
Total	-0.84	-0.71	-0.13	-2.5*	-0.51	-0.39	-0.12	-7.1**	-0.11	-0.04	-0.07	-8.2**
Air Force												
E	0.30	-0.38	0.68	1.4	0.18	-0.15	0.33	1.3	0.21	0.13	0.07	0.7
M	-0.26	-0.09	-0.17	-0.3	-0.17	-0.02	-0.15	-0.6	0.05	0.10	-0.05	-0.7
Army												
EL	-0.94	-0.57	-0.38	-3.3**	-0.53	-0.28	-0.25	-6.5**	-0.15	0.02	-0.17	-10.6**
GM	-0.08	-0.57	0.48	2.9**	-0.13	-0.28	0.16	3.0**	0.11	-0.07	0.18	5.8**
MM	-1.59	-1.01	-0.58	-5.7**	-0.91	-0.50	-0.41	-13.5**	-0.34	-0.03	-0.30	-14.6**
OF	-0.89	-0.88	-0.01	-0.1	-0.52	-0.50	-0.03	-1.1	-0.16	-0.15	-0.01	-0.5
SC	-0.78	-0.57	-0.21	-1.6	-0.36	-0.34	-0.02	-0.4	-0.02	-0.09	0.06	2.6**
Navy												
EL	0.32	0.21	0.11	0.2	0.08	0.08	0.01	0.0	0.08	0.13	-0.05	-0.7
EG	0.31	-0.42	0.73	1.5	-0.11	-0.29	0.17	1.0	-0.23	-0.10	-0.13	-2.8**

Comp.	Prediction at +0.5 s.d. Fem.	Male	Diff	t	Prediction at +1.0 s.d. Fem.	Male	Diff	t
Total	0.28	0.29	-0.02	-2.0*	0.64	0.63	0.01	1.1
Air Force								
E	0.46	0.48	-0.03	-0.9	0.79	0.86	-0.08	-3.5**
M	0.34	0.30	0.04	1.0	0.75	0.55	0.21	2.9**
Army								
EL	0.23	0.30	-0.07	-3.5**	0.55	0.58	-0.03	-1.3
GM	0.39	0.23	0.16	4.6**	0.77	0.55	0.22	4.0**
MM	0.12	0.34	-0.23	-8.2**	0.44	0.65	-0.21	-3.8**
OF	0.23	0.20	0.02	1.4	0.66	0.54	0.12	3.3**
SC	0.31	0.19	0.12	5.2**	0.66	0.50	0.16	4.4**
Navy								
EL	0.32	0.42	-0.09	-2.5*	0.72	0.75	-0.03	-1.0
EG	-0.03	0.15	-0.19	-2.8**	0.50	0.48	0.02	0.2

* - difference significant at the .05 (two-tail) level;
** - difference significant at the .01 (two-tail) level

The results by race indicate that, for each predictor score level, whites had significantly higher expected criterion scores. While the differences are of statistical significance in these very large samples, they are of somewhat limited practical significance, being only about one-tenth of a standard deviation. (With this size difference, for example, roughly 46% of the blacks at a selection score level will score above the criterion mean for whites at that level.) Most of the individual composites also showed significant overprediction for blacks. The only significant differences in the opposite direction were found for the Army SC composite.

The overall results by sex were quite similar to the results by race, with males having significantly higher criterion scores at all but the highest level of the selection test scale. In these analyses, the Army GM and SC composites both showed results counter to the overall trend at several points in the range of interest. Again the size of the differences is quite small, notwithstanding the statistical significance in these large samples. At the high end of the scale, the area of greatest interest in the GAO's analyses, the average differences are literally zero.

Marine Corps Job Performance Measurement Project

The analyses of the Marine Corps Job Performance Measurement Project proceeded somewhat differently from the analyses reported here. In particular, those data were collected for research only, while the data reported above used operational scores for each recruit, so greater attention was given to eliminating outliers that might reflect lack of motivation or other factors associated with research-only data. Nonetheless, the results of the Marine Corps analyses were entirely consistent with the above fmdings. The difference in regression slopes between blacks and whites was not significant. The difference between the regression lines was also not significant but in the same direction as the aggregate results in the present study. The data used in this analyses were not available for pooling with results from the other data sets, but the sample size, 118 blacks and 632 whites, was too small to have had any significant effect on the overall results. Appendix A contains more information on analyses of the Marine Corps data.

Conclusions

The general conclusion from the analyses is that the ASVAB technical composites are highly sensitive predictors of training and job performance for all applicant groups. Contrary to the GAO's findings, these composites were found to be more sensitive predictors for females than for males. Small but significant differences indicating greater sensitivity for whites than for blacks do suggest the need for further investigation and possible refmements in the battery and the technical composites derived from the battery.

The small but persistent differences in the prediction functions suggest that there are other characteristics, not measured by the current ASVAB, which are related to job outcomes and on which the applicant groups differ. As new measures are considered for inclusion in the ASVAB, it will be important to evaluate the extent to which such differences might be accounted for.

Overall, the results do not suggest the need for urgent changes in the current ASVAB or in the selection composites derived from the ASVAB. Nonetheless, proposed changes are currently under evaluation. New measures under consideration include spatial, psychomotor, and memory tests. It is possible, but by no means certain, that the characteristics measured by these new tests will be less related to the opportunity to learn. Consequently, there may be smaller differences among applicant groups in these new tests in comparison to many of the tests in the current battery. The impact of these new measures on the sensitivity and fairness of the battery as a whole will be carefully evaluated in deciding whether they should be used operationally.

In addition to considering new measures, the Services continue to review their selection composites and to consider changes. The analyses reported here provide a model for investigation of the sensitivity and fairness of any new composites for all applicant groups.

REFERENCES

American Educational Research Association, American Psychological Society, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Bock, R. D., & Moore, E. G. (1984). *Profile of American youth: demographic influences in ASVM test performance.* Washington, DC: Office of the Assistant Secretary of Defense (Manpower, Installations, and Logistics).

Booth-Kewley, S., Foley, P. P., & Swanson, L. (1984). *Predictive validation of the Armed Services Vocational Aptitude Battery (ASVM) Forms 8, 9, and 10 against 100 Navy schools.* (NPRDC-TR-85-15). San Diego, CA: Navy Personnel Research and Development Center.

Cleary, T. A. (1968). Test bias: Prediction of grades for Negro and white students in integrated colleges. *Journal of Educational Measurement, 5*, 115-124.

Department of Defense. (1984a). *Armed Services Vocational Aptitude Battery (ASVAB) test manual.* (DoD 1304.12AA). Chicago, IL: Military Entrance Processing Command.

Department of Defense. (1984b). *Technical supplement to the counselor's manual for ASVM 14.* North Chicago, IL: Military Entrance Processing Command.

Eitelberg, M. J. (1988). *Manpower for military occupations.* Alexandria, VA: Human Resources Research Organization.

Eitelberg, M. J., Laurence, J. H., Waters, B. K., & Perelman, L. S. (1984). *Screening for service: aptitude and education criteria for military entry.* Alexandria, VA: Human Resources Research Organization.

General Accounting Office (October 16, 1990). *Military training: its effectiveness for technical specialties is unknown.* (GAO code 973276) OSD Case 8371.

Gifford, B. R. (Ed.). (1989). *Test policy and the politics of opportunity allocation: the workplace and the law.* Boston, MA: Kluwer Academic Press.

Hartigan, J. A., & Widgor, A. K. (Eds.). (1989). *Fairness in employment testing: validity generalization, minority issues, and the General Aptitude Test Battery.* Washington, DC: National Academy Press.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Hunter, J. E. (1983). *Fairness of the General Aptitude Test Battery (GATB): ability differences and their impact on minority hiring rates* (Test Research Report No. 46). Washington, DC: Employment Service, U.S. Department of Labor.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: correcting error and bias in research findings*. Newbury Park, CA: Sage Publications, Inc.

Linn, R. L. (1982). Ability testing: individual differences, prediction and differential prediction. In A. K. Wigdor & W. R. Garner (Eds.), *Ability testing: uses, consequences, and controversies, part II*. Washington, DC: National Academy Press.

Linn, R. L., & Dunbar, S. B. (1986). Validity generalization and predictive bias. In R. A. Berk (Ed.), *Performance assessment: methods & applications*. Baltimore, MD: The Johns Hopkins University Press.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley Publishing Company.

Maier, M. H., & Hirshfeld, S. F. (1978). *Criterion-referenced job proficiency testing: a large scale application*. ARI Research Report No. 1193. Alexandria, VA: Army Research Institute for the Behavioral Sciences.

Maier, M., & Truss, A. R. (1984). *Validity of the occupational and academic composites for the Armed Services Vocational Aptitude Battery (ASVAB), form 14, in the Marine Corps training courses*. (CNA Tech. Rept. 84-3043/2). Alexandria, VA: Center for Naval Analyses.

McLaughlin, D. H., Rossmeissl, P. G., Wise, L. L., Brandt, D. A., & Wang, M. (1984). *Validation of current alternative Armed Services Vocational Aptitude Battery (ASVAB) area composites, based on training and Skill Qualification Test (SQT) information in fiscal year 1981 and 1982*. (ARI-TR-651, AD A156807). Alexandria, VA: Army Research Institute.

Ree, M. J., & Earles, J. A. (1990). *Differential validity of a differential aptitude test*. (AFHRL-TR-89-59). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Welsh, J. R., Kucinkas, S. K., & Curran, L. T. (1990). *Armed Services Vocational Aptitude Battery (ASVAB): integrative review of validity studies*. (AFHRL-TR-90-). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Weltin, M. M., & Popelka, B. A. (1983). *Evaluation of the ASVAB 8, 9, and 10 clerical composite for predicting training school performance*. (ARI Tech. Rep. No. 594). Alexandria, VA: Army Research Institute.

Wilbourn, J. M., Valentine, L. D., & Ree, M. J. (1984). *Relationships of the Armed Services Vocational Aptitude Battery (ASVAB) forms 8, 9, and 10 to the Air Force technical school grades*. (AFHRL-TR-84-8, AD-A144-213). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

APPENDIXES

Appendix A

Subgroup Effects in the Prediction of Hands-on Performance Scores for the Marine Corps Automotive Mechanic Specialty

To investigate sensitivity and fairness of the ASVAB technical composites in the Marine Corps, several factors were studied:

- the Marine Corps hands-on performance test (HOPT) for the Automotive Mechanic specialty;
- time in service (TIS);
- enlistment ASVAB composites; and
- current computer-adaptive ASVAB composites (CAT-ASVAB).

Discussion follows.

In its Job Performance Measurement (JPM) project, the Marine Corps developed a hands-on performance test (HOPT) for the Automotive Mechanic specialty (MOS 3521). The test consists of a sample of tasks that a mechanic needs to perform in the course of his or her work. Each task was divided into a number of steps; each step was scored as performed correctly or not. The test was administered by former Marines who had relevant job experience and were trained to score performance objectively. Wigdor and Green (1986, p. 95) refer to such a score as the "benchmark measure" of job performance.

Time in service (TIS) has been found to be a powerful predictor of hands-on performance. Given equal ASVAB scores, senior Marines score higher on the HOPT, on the average, than junior Marines. This increase results from training on the job. The rate of growth slows as time increases (note exclusions below). Therefore, TIS and its square were included as predictors, along with the ASVAB scores.

The available ASVAB technical composites were those the Marine enlisted with, plus composites from a computer-adaptive version of the ASVAB (CAT-ASVAB) that was administered the day after the HOW. Occupational composites used by the Marine Corps have a mean of 100 and a standard deviation of 20 in the national population. The composite used for the Automotive Mechanic occupation is Mechanical Maintenance (MM).

The MM composite is considered fair to black males if the regression of the HOPT on the MM is the same for black males as for white males. Standard statistical tests were performed using a Statistical Analysis System (SAS) program. Equal slopes in the two groups imply that the MM composite is equally sensitive for both groups. Equal intercepts imply that there is no over- or underprediction from the HOPT for either group.

One problem was that the minority sample size was originally only 118, much smaller than the minimum of 400 per composite used in analyzing data from the other Services. When sample size is small, a few highly influential cases can change the result substantially. Therefore each significance test was preceded by influence analysis. Cases with extreme values of the influence function were excluded, and then a significance test was performed on the edited sample.

Excluded from the study were

- females and Hispanics, because their numbers were too small for useful analysis;
- Marines whose TIS exceeded ten years (4 cases);
- cases with extreme values of influence (12 cases).

The remaining sample, with complete data for each Marine, contained 106 black males and 632 white males.

In the influence analysis of the MM composite obtained at time of enlistment, the regression equation initially included a term to represent the difference in slopes between black males and white males. Influence on this term was calculated for all individuals in the sample. The standard deviation of the influence values was ,038, while the mean was zero, as theory requires. Using the edited sample, the F ratio for difference between slopes was 0.54, which is statistically nonsignificant. Therefore, in the analysis of difference between intercepts, slopes in the two groups were set to be equal. Then influence analysis was performed for difference between intercepts. Standard deviation of influence values was .041. Again, cases with influence above .25 in magnitude were deleted. This further reduced the sample size by three. The F ratio for difference between intercepts was 3.62, which is not significant at the .05 level.

A similar procedure was followed with the MM composite obtained from the CAT-ASVAB. The cutoff value for size of influence was again .25. Three cases were deleted for the analysis of slopes and two more for the analysis of intercepts.

Regression coefficients, F ratios, and tail probabilities using the enlistment ASVAB and the CAT-ASVAB composites were as follows:

	Enlistment ASVAB			CAT-ASVAB		
	Black Males		White Males	Black Males		White Males
Slope						
Estimates	.22		.31	.38		.35
F ratio		0.54			0.17	
Significance level		.46			.68	
Intercept						
Estimates	37.67		39.15	32.70		34.09
F ratio		3.62			3.58	
Significance level		.057			.059	

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

In summary, the Marine Corps JPM results for the Automotive Mechanic specialty, using the hands-on performance test as the criterion, show that the MM composite is equally sensitive for both black and white males. The results also show that the regression equation does not over- or underpredict the performance of black males.

Appendix B

Sample Sizes for Navy Schools Used in the Analyses*

CDP/RATING		DESCRIPTION	B	W	F	M
		EL: Electronics Composite				
6501	AD	Aviation Mechanic	355	2375	229	2714
6506	AO	Aviat. Ordnanceman	138	559		
6240	AQ	Aviat.Fire Contrl.Tech.	59	871		
6239	AT	Aviat. Elect. Tech.	166	3560	399	3413
6241	AX	Aviat. Elect. Tech	43	911		
6161	CTM	Cryptolog. Tech. Maint.			87	268
6131	DS	Data Sytems Tech.			41	207
615L	ET1	Electronics Tech.(ph 1)	295	2809	441	2733
603V	ET2	Electronics Tech.(ph 2)	194	2124	337	2046
609W	FC	Fire Control Tech.	79	1691		
6400	GM	Gunner's Mate	49	371		
611T	IC/4YO	Interior Com. Tech.	247	1168	160	1292
6015	STG	Sonar Technician	129	1666		
		EG: Engineering Composite				
6612	BT/4YO	Boiler Technician	481	3153	40	3805
6613	BT/6YO	Boiler Technician	50	635		
6487	EN/4YO	Engineman	368	3167	338	3385
6611	MM/4YO	Machinists Mate	645	4141	88	4996
		ME: Mechanical Composite				
6097	EO	Equipment Operator	53	663		
6519	PR	AirCrw. Survl. Equipmn.			41	372
		MR: Machinery Repair				
6513	ABE	Avait Btwsns Mate (EQP)	99	290		
6512	ABF	Avait Btwsns Mate (FLS)	86	239		
6517	ABH	Avait. Str.Mech (Hydrl)	130	426		
6068	MR	Machinery Repairman	54	915	158	828

* CDP = Course Data Processing Number, Rating indicated job code.

Appendix C

Sample Sizes for Air Force Apprentice-level Specialties Used in the Analyses*

Cmp	AFSC	Sample Sizes Description	B	W	F	M
G	12230	Acrw Life Suprt Spec	187	313	85	342
G	20130	Intel Ops Spec	147	320	114	263
G	20230	Radio Com Analy Specl	48	373	136	302
G	20630	Imagry Interprtr Specl			66	168
A	20731	Morse Sys Oper	138	352	151	355
G	20833	Crypto Ling Specl			65	335
G	23330	Imagery Prod Specl	38	178	77	154
G	25130	Weather Specl	61	519	165	457
A	27132	Ops Resource Mgt Specl			82	92
G	27230	Air Traffic Ctrl Opr	156	880	280	805
G	27430	Command and Ctrl Specl	70	251	108	234
G	27630	Aerospace Con & Warn Sys Opr			42	77
G	27630B	" " 416L SAGE			43	77
G	27630C	" " 407L TACS	115	523	162	510
E	30430	Wideband Com Eqp Specl			53	195
E	30431	Nav Aid Equip Specl			41	183
E	30434	Grnd Radio Equip Specl	152	1218	277	1178
E	30630	Elect Comp&Crypto Eq Specl	40	321	73	305
E	30633	Telecorn Sys Maint Specl	32	250	61	242
E	32430	Prec Msmt Equip Lab Specl	50	647	118	620
E	32530	Avionics Flgt Contr Specl	50	410	92	387
E	32531	Avionics Instr Sys Specl	64	555	111	538
E	32830	Avionics Com Sys Specl	57	523	104	507
E	32831	Avionics Nav Sys Specl	58	523	87	525
E	32833	Elect Warfare Sys Specl	46	446		
E	36231	Telephone switching Specl			42	159
G	39130	Maint Data Syst Analy Tech	47	148	58	156
G	39230	Maintenance Schedul Specl	97	354	148	343
E	41130B	" " BGM-109	44	332	74	330
M	41131A	Msl Maint Specl WS-133	47	476		
E	41132A	Msl Facilts Specl WS1338	42	219		
E	42330	Acrft Elect Sys Specl	133	704	100	803
M	42331	Acrft Env Sys Specl	63	319	63	334
M	42731	Corrosive Cont Specl			53	365
M	42735	Air Frame Repr Specl	44	686	57	707
E	45234	Tac Acrft Maint Specl	106	1610	98	1683
E	45430A	Aerosp Proplsn Specl JE	52	726		
M	45433	Acft Fuel Sys Specl	43	531	64	527
E	45434	Acft Pneudraulic Sys Spc	45	423		
E	45730	Bomb-Nav Sys Specl	59	1103	98	1104
E	45732	Airlift Acft Maint Specl	57	876	50	922
G	45831	Non Destr Inspect Specl	43	160	68	146
M	46130	Munitions Sys Specl	127	1606	106	1696
M	46230F	" " " " F-16	48	610		
A	46530	Munitions Ops Specl	66	147	77	149
G	49131	Com - Comp Sys Opr	120	1989	509	1666
G	49132	Com-Comp Sys Progrm Specl			53	218
A	49231	Com Sys Radio Oper	173	367	215	373
E	49330	Com Sys Electrng Spect Mgt	66	434	101	432

continued

Appendix C
(continued)

Sample Sizes for Air Force Apprentice-level Specialties Used in the Analyses*

Cmp	AFSC	Sample Sizes Description	B	W	F	M
G	49630	Com-Comp Sys P & P Mgt Spc			53	130
M	54232	Elect Powr Prod Specl	41	333		
G	55330	Engineering Asst Specl	44	137	46	162
G	55530	Production Contrl Specl			55	74
M	56631	Environ Support Specl	51	199		
G	57130	Fire Protection Specl	269	1687		
G	60130	Packing Specl	60	227	67	248
A	60230	Passngr 7 HHG Specl	84	165	91	177
A	60231	Freight & Pkgng Specl	100	252	144	255
A	60530	Air Passenger Specl	75	225	152	175
M	60531	Air Cargo Specl	74	605	82	653
G	62330	Services Specl	188	543	226	585
M	63130	Fuel Specl	136	1417	93	1535
A	64530	Inventory Mgmt Specl	952	2164	1166	2261
G	64531	Mat Strg & Distr Specl	117	214	115	252
A	67231	Financial Mgmt Specl	117	334	202	285
A	67232	Financial Services Specl	191	462	310	408
A	70130	Chapel Mgmt Specl			66	69
A	70230	Information Mgmt Specl	1254	2248	1406	2483
A	73230	Career Advisory Specl	484	1090	816	900
A	73231	Personal Affairs Specl	40	67	63	54
G	81130	Security Specl	1113	6881	878	7478
G	81132	Law Enforcement Specl	624	3234	751	3316
G	81132A	Law Enf Working Dog Qual	45	485	95	455
G	81150	Security Specl	87	554	43	628
G	90130	Aeromedical Specl	70	204	109	198
G	90230	Medical Services Specl	431	1602	859	1385
G	90232	Surgical Services Specl	43	150	101	120
G	90330	Radiologic Specl	54	240	107	212
G	90530	Pharmacy Specl	66	191	111	170
G	90630	Medical Admin Specl	215	620	324	597
G	90730	Bioeng Specl	54	146	83	140
G	90830	Environmental Medcn Specl	46	142	65	145
G	91330	Physical Therapy Specl			45	81
G	91530	Medical Material Specl	93	252	150	227
G	92430	Medical Lab Specl	140	326	176	350
G	92630	Diet Therapy Specl	47	172	100	136
G	98130	Dental Assist Specl	158	566	295	505
G	98230	Dental Lab Specl			62	159

*Cmp indicates selection composite; AFSC is Air Force Specialty Code

Appendix D

Sample Sizes for Army Specialties Used in the Analyses By Selection Composite*

MOS	Year	Description	B	W	F	M	Prior	New
		Electronics (EL) Composite						
26Q	85	TACTICAL SATELLITE/MICROWAVE SYSTEM OPER	109	364	51	435	26Q	31Q
26Q	86	TACTICAL SATELLITE/MICROWAVE SYSTEM OPER	160	423	60	550	26Q	31Q
26Q	87	TACTICAL SATELLITE/MICROWAVE SYSTEM OPER	159	372	46	521	26Q	31Q
26V	86	STRATEGIC MICROWAVE SYSTEMS REPAIRER			54	322	26V	29V
27E	85	TOW/DRAGON REPAIRER	89	275			27H	27E
27E	86	TOW/DRAGON REPAIRER	149	363			27H	27E
27E	87	TOW/DRAGON REPAIRER	144	311			27H	27E
27E	88	TOW/DRAGON REPAIRER	114	239			27H	27E
27E	89	TOW/DRAGON REPAIRER	105	189			27H	27E
29E	88	RADIO REPAIRER	74	443			31E	29E
29E	89	RADIO REPAIRER	85	519	.	.	31E	29E
29J	86	TELECOMMUNICATIONS TERMINAL DEVICE REPAI	72	437	41	486	31J	29J
29J	87	TELECOMMUNICATIONS TERMINAL DEVICE REPAI	71	335	.	.	31J	29J
29J	88	TELECOMMUNICATIONS TERMINAL DEVICE REPAI	61	354	.	.	31J	29J
29N	88	TELEPHONE CENTRAL OFFICE REPAIRER	145	291	.	.	36H	29N
29N	89	TELEPHONE CENTRAL OFFICE REPAIRER	124	242	.	.	36H	29N
29V	87	STRATEGIC MICROWAVE SYSTEMS REPAIRER	41	323	42	338	26R	29V
29V	89	STRATEGIC MICROWAVE SYSTEMS REPAIRER	44	399	.	.	26R	29V
31J	85	TELECOMMUNICATIONS TERMINAL DEVICE REPAI	51	401	49	415	31J	29J
31K	86	COMBAT SIGNALER	1059	1631	171	2660	05B	31K
31K	87	COMBAT SIGNALER	1114	1671	239	2708	05B	31K
31K	88	COMBAT SIGNALER	1343	2034	265	3312	05B	31K
31K	89	COMBAT SIGNALER	137	137	.	.	05B	31K
31L	88	WIRE SYSTEMS INSTALLER	768	795	332	1316	36C	31L
31L	89	WIRE SYSTEMS INSTALLER	538	520	208	909	36C	31L
31M	85	MULTICHANNEL COMMUNICATIONS SYSTEMS OPER	790	1757	386	2251		31M
31M	86	MULTICHANNEL COMMUNICATIONS SYSTEMS OPER	838	1757	476	2231		31M
31M	87	MULTICHANNEL COMMUNICATIONS SYSTEMS OPER	928	1844	447	2442		31M
31M	88	MULTICHANNEL COMMUNICATIONS SYSTEMS OPER	1277	2373	445	3379		31M
31M	89	MULTICHANNEL COMMUNICATIONS SYSTEMS OPER	1252	1988	443	2955		31M
31M	90	MULTICHANNEL COMMUNICATIONS SYSTEMS OPER	529	688	152	1117		31M
31N	85	COMMUNICATIONS SYSTEMS/CIRCUIT CONTROLLE	82	153	60	187	32D	31N
31N	86	COMMUNICATIONS SYSTEMS/CIRCUIT CONTROLLE	71	113	73	124	32D	31N
31N	87	COMMUNICATIONS SYSTEMS/CIRCUIT CONTROLLE	74	82	74	88	32D	31N
31N	88	COMMUNICATIONS SYSTEMS/CIRCUIT CONTROLLE	74	88	71	99	32D	31N
31N	89	COMMUNICATIONS SYSTEMS/CIRCUIT CONTROLLE	104	108	63	159	32D	31N
31Q	88	TACTICAL SATELLITE/MICROWAVE SYSTEM OPER	155	395	46	537	26Q	31Q
31Q	89	TACTICAL SATELLITE/MICROWAVE SYSTEM OPER	168	457	66	591	26Q	31Q
31Q	90	TACTICAL SATELLITE/MICROWAVE SYSTEM OPER	107	250	.	.	26Q	31Q
31V	85	UNIT LEVEL COMMUNICATIONS MAINTAINER	228	732	41	950	31G	31G
31V	86	UNIT LEVEL COMMUNICATIONS MAINTAINER	398	1219	80	1621	31G	31G
31V	87	UNIT LEVEL COMMUNICATIONS MAINTAINER	586	1617	122	2181	31G	31G
31V	88	UNIT LEVEL COMMUNICATIONS MAINTAINER	601	1537	147	2114	31G	31G
31V	89	UNIT LEVEL COMMUNICATIONS MAINTAINER	406	832	92	1219	31G	31G
32D	85	COMMUNICATIONS SYSTEMS/CIRCUIT CONTROLLE	116	362	114	385	32D	31N
32D	86	COMMUNICATIONS SYSTEMS/CIRCUIT CONTROLLE	125	349	128	366	32D	31N
32D	87	COMMUNICATIONS SYSTEMS/CIRCUIT CONTROLLE	150	444	177	440	32D	31N
32D	88	COMMUNICATIONS SYSTEMS/CIRCUIT CONTROLLE	181	419	169	456	32D	31N
35K	87	AVIONIC MECHANIC	58	198	47	217	35K	68N
35K	88	AVIONIC MECHANIC	57	207	44	236	35K	68N
35K	89	AVIONIC MECHANIC	89	366	61	419	35K	68N
35L	89	AVIONIC COMMUNICATIONS EQUIPMENT REPAIRE	53	158	.	.	35L	68L
36C	86	WIRE SYSTEMS INSTALLER	991	1108	497	1740	36C	31L
36C	87	WIRE SYSTEMS INSTALLER	784	776	343	1305	36C	31L

continued

Appendix D
(continued)

Sample Sizes for Army Specialties Used in the Analyses By Selection Composite*

MOS	Year	Description	B	U	F	M	Prior	New
36H	85	TELEPHONE CENTRAL OFFICE REPAIRER	42	161			36H	29N
36M	85	SWITCHING SYSTEMS OPERATOR	44	114	.	.		36M
36M	86	SWITCHING SYSTEMS OPERATOR	116	235	115	258		36M
36M	87	SWITCHING SYSTEMS OPERATOR	234	367	192	437		36M
36M	88	SWITCHING SYSTEMS OPERATOR	266	372	210	453		36M
51R	86	INTERIOR ELECTRICIAN	46	283	.			51R
51R	87	INTERIOR ELECTRICIAN	74	357	.			51R
51R	88	INTERIOR ELECTRICIAN	58	291	.			51R
51R	89	INTERIOR ELECTRICIAN	40	173	.	.		51R
55G	85	NUCLEAR WEAPONS SPECIALIST	.	.	42	94	35F	55G
68J	88	AIRCRAFT ARMAMENT/MISSILE SYSTEMS REPAIR	54	291	.	.		68J
68J	89	AIRCRAFT ARMAMENT/MISSILE SYSTEMS REPAIR	85	492	.	.		68J
68J	90	AIRCRAFT ARMAMENT/MISSILE SYSTEMS REPAIR	48	261	.	.		68J
96R	85	GROUND SURVEILLANCE SYSTEMS OPERATOR	105	329	.	.	17K	96R
96R	86	GROUND SURVEILLANCE SYSTEMS OPERATOR	71	236	.	.	17K	96R
96R	87	GROUND SURVEILLANCE SYSTEMS OPERATOR	42	124	.	.	17K	96R
96R	88	GROUND SURVEILLANCE SYSTEMS OPERATOR	58	203	.	.	17K	96R
96R	89	GROUND SURVEILLANCE SYSTEMS OPERATOR	49	171	.	.	17K	96R

General Maintenance (GM) Composite

MOS	Year	Description	B	U	F	M	Prior	New
41C	85	FIRE CONTROL INSTRUMENT REPAIRER	46	106				41C
41C	88	FIRE CONTROL INSTRUMENT REPAIRER	45	50	.			41C
42D	85	DENTAL LABORATORY SPECIALIST		.	46	80		42D
43E	85	PARACHUTE RIGGER	97	403	67	468		43E
43E	86	PARACHUTE RIGGER	95	406	79	454		43E
43E	87	PARACHUTE RIGGER	80	349	59	393		43E
43E	88	PARACHUTE RIGGER	84	421	.			43E
43E	89	PARACHUTE RIGGER	111	432	.			43E
43M	85	FABRIC REPAIR SPECIALIST	76	41	.			43M
43M	86	FABRIC REPAIR SPECIALIST	90	41	.			43M
44B	85	METAL WORKER	56	297	.			44B
44B	86	METAL WORKER	99	447	.			44B
44B	87	METAL WORKER	127	507	.			44B
44B	88	METAL WORKER	130	485	.			44B
44B	89	METAL WORKER	92	292	.			44B
45B	88	SMALL ARMS REPAIRER	43	214	.			45B
45B	89	SMALL ARMS REPAIRER	41	195	.			45B
45K	85	TANK TURRET REPAIRER	45	258	.			45K
45K	86	TANK TURRET REPAIRER	53	302	.			45K
45K	87	TANK TURRET REPAIRER	67	289	.			45K
45K	88	TANK TURRET REPAIRER	73	347	.			45K
45K	89	TANK TURRET REPAIRER	51	245	.			45K
45T	87	BRADLEY FIGHTING VEHICLE SYSTEM TURRET M	50	160	.			45T
45T	88	BRADLEY FIGHTING VEHICLE SYSTEM TURRET M	53	164	.			45T
45T	89	BRADLEY FIGHTING VEHICLE SYSTEM TURRET M	49	145	.			45T
51B	85	CARPENTRY AND MASONRY SPECIALIST	104	395	.		51C	51B
51B	86	CARPENTRY AND MASONRY SPECIALIST	126	416	.		51C	51B
51B	87	CARPENTRY AND MASONRY SPECIALIST	170	596	.		51C	51B
51B	88	CARPENTRY AND MASONRY SPECIALIST	247	921	.		51C	51B
51B	89	CARPENTRY AND MASONRY SPECIALIST	213	712	.		51C	51B
51K	85	PLUMBER	98	169	.			51K

continued

Appendix D
(continued)

Sample Sizes for Army Specialties Used in the Analyses By Selection Composite*

MOS	Year	Description	B	W	F	M	Prior	New
51K	86	PLUMBER	95	167	.	.		51K
51K	87	PLUMBER	98	172	.	.		51K
51K	88	PLUMBER	97	186	.	.		51K
51K	89	PLUMBER	72	117	.	.		51K
51N	85	WATER TREATMENT SPECIALIST	84	96	.	.	51N	77W
51N	86	WATER TREATMENT SPECIALIST	138	129	.	.	51N	77W
52C	86	UTILITY EQUIPMENT REPAIRER	81	454	61	491		52X
52C	89	UTILITY EQUIPMENT REPAIRER	155	603	80	712		52X
52D	85	POWER GENERATOR EQUIPMENT REPAIRER	332	1360	90	1679		52D
52D	86	POWER GENERATOR EQUIPMENT REPAIRER	507	1896	131	2431		52D
52D	87	POWER GENERATOR EQUIPMENT REPAIRER	504	2025	105	2587		52D
52D	88	POWER GENERATOR EQUIPMENT REPAIRER	681	2467	139	3216		52D
52D	89	POWER GENERATOR EQUIPMENT REPAIRER	521	1615	72	2216		52D
55B	85	AMMUNITIONS SPECIALIST	223	497	73	684		55B
55B	86	AMMUNITIONS SPECIALIST	318	602	68	901		55B
55B	87	AMMUNITIONS SPECIALIST	323	739	74	1041		55B
55B	88	AMMUNITIONS SPECIALIST	357	824	106	1134		55B
55B	89	AMMUNITIONS SPECIALIST	300	656	86	916		55B
55B	90	AMMUNITIONS SPECIALIST	102	204	.	.		55B
57E	85	LAUNDRY AND BATH SPECIALIST	145	78	60	177		57E
57E	86	LAUNDRY AND BATH SPECIALIST	162	88	72	196		57E
57E	87	LAUNDRY AND BATH SPECIALIST	149	66	48	182		57E
57E	88	LAUNDRY AND BATH SPECIALIST	148	68	45	189		57E
57E	89	LAUNDRY AND BATH SPECIALIST	104	55	.	.		57E
57F	89	GRAVES REGISTRATION SPECIALIST	40	70	.	.		57F
57H	85	CARGO SPECIALIST	225	306	110	441	57H	88H
57H	86	CARGO SPECIALIST	252	280	93	464	57H	88H
57H	87	CARGO SPECIALIST	302	411	73	672	57H	88H
62E	85	HEAVY CONSTRUCTION EQUIPHENT OPERATOR	114	803	.	.		62E
62E	86	HEAVY CONSTRUCTION EQUIPMENT OPERATOR	117	792	.	.		62E
62E	87	HEAVY CONSTRUCTION EQUIPMENT OPERATOR	125	795	.	.		62E
62E	88	HEAVY CONSTRUCTION EQUIPMENT OPERATOR	146	886	.	.		62E
62E	89	HEAVY CONSTRUCTION EQUIPMENT OPERATOR	180	1040	.	.		62E
62F	85	CRANE OPERATOR	78	275	.	.		62F
62F	86	CRANE OPERATOR	71	248	.	.		62F
62F	87	CRANE OPERATOR	76	250	.	.		62F
62F	88	CRANE OPERATOR	64	232				62F
62F	89	CRANE OPERATOR	51	163	.	.		62F
62J	85	GENERAL CONSTRUCTION EQUIPMENT OPERATOR	109	422	.	.		62J
62J	86	GENERAL CONSTRUCTION EQUIPMENT OPERATOR	126	516	.	.		62J
62J	87	GENERAL CONSTRUCTION EQUIPMENT OPERATOR	118	466	.	.		62J
62J	88	GENERAL CONSTRUCTION EQUIPHENT OPERATOR	104	423	.	.		62J
62J	89	GENERAL CONSTRUCTION EQUIPMENT OPERATOR	142	512	.	.		62J
77W	87	WATER TREATMENT SPECIALIST	152	125	.	.	51N	77W
77W	88	WATER TREATMENT SPECIALIST	154	146	54	267	51N	77W
77W	89	WATER TREATMENT SPECIALIST	127	117	.	.	51N	77W
88H	88	CARGO SPECIALIST	249	362	63	576	57H	88H
88H	89	CARGO SPECIALIST	230	267	54	465	57H	88H

continued

Appendix D
(continued)

Sample Sizes for Army Specialties Used in the Analyses By Selection Composite*

Mechanical Maintenance (MM) Composite

MOS	Year	Description	B	W	F	M	Prior	New
45E	87	M1 ABRAMS TANK TURRET MECHANIC	58	165	.	.		45E
45E	88	M1 ABRAMS TANK TURRET MECHANIC	67	233	.	.		45E
45E	89	M1 ABRAMS TANK TURRET MECHANIC	59	196	.	.		45E
45N	85	M60A1/A3 TANK TURRET MECHANIC	55	233	.	.		45N
45N	86	M60A1/A3 TANK TURRET MECHANIC	60	223	.	.		45N
45N	87	M60A1/A3 TANK TURRET MECHANIC	61	166	.	.		45N
62B	85	CONSTRUCTION EQUIPMENT REPAIRER	221	797	44	1014		62B
62B	86	CONSTRUCTION EQUIPMENT REPAIRER	290	1061	65	1336		62B
62B	87	CONSTRUCTION EQUIPMENT REPAIRER	349	1229	55	1595		62B
62B	88	CONSTRUCTION EQUIPMENT REPAIRER	407	1271	51	1717		62B
62B	89	CONSTRUCTION EQUIPMENT REPAIRER	357	1023	.	.		62B
62B	90	CONSTRUCTION EQUIPMENT REPAIRER	86	212	.	.		62B
63B	85	LIGHT-WHEEL VEHICLE MECHANIC	1634	4508	443	6034		63B
63B	86	LIGHT-WHEEL VEHICLE MECHANIC	1665	4914	494	6422		63B
63B	87	LIGHT-WHEEL VEHICLE MECHANIC	1659	5075	518	6549		63B
63B	88	LIGHT-WHEEL VEHICLE MECHANIC	1681	4585	539	6061		63B
63B	89	LIGHT-WHEEL VEHICLE MECHANIC	1371	3541	417	4756		63B
63D	86	SELF-PROPELLED FIELD ARTILLERY SYSTEM ME	41	495	.	.		63D
63D	87	SELF-PROPELLED FIELD ARTILLERY SYSTEM ME	51	527	.	.		63D
63D	88	SELF-PROPELLED FIELD ARTILLERY SYSTEM ME	58	588	.	.		63D
63D	89	SELF-PROPELLED FIELD ARTILLERY SYSTEM ME	42	410	.	.		63D
63E	86	M1 ABRAMS TANK SYSTEM MECHANIC	56	356	.	.		63E
63E	87	M1 ABRAMS TANK SYSTEM MECHANIC	63	517	.	.		63E
63E	88	M1 ABRAMS TANK SYSTEM MECHANIC	104	629	.	.		63E
63E	89	M1 ABRAMS TANK SYSTEM MECHANIC	94	503	.	.		63E
63G	86	FUEL AND ELECTRICAL SYSTEM REPAIRER	48	273	.	.		63G
63G	87	FUEL AND ELECTRICAL SYSTEM REPAIRER	60	360	.	.		63G
63G	88	FUEL AND ELECTRICAL SYSTEM REPAIRER	87	388	.	.		63G
63G	89	FUEL AND ELECTRICAL SYSTEM REPAIRER	61	270	.	.		63G
63H	85	TRACK VEHICLE REPAIRER	297	1037	.	.		63H
63H	86	TRACK VEHICLE REPAIRER	233	812	56	1030		63H
63H	87	TRACK VEHICLE REPAIRER	321	930	70	1230		63H
63H	88	TRACK VEHICLE REPAIRER	399	956	93	1322		63H
63H	89	TRACK VEHICLE REPAIRER	322	604	55	921		63H
63J	85	QUARTERMASTER AND CHEMICAL EQUIPMENT REP	146	169	.	.	54D	63J
63J	86	QUARTERMASTER AND CHEMICAL EQUIPMENT REP	166	223	62	343	54D	63J
63J	87	QUARTERMASTER AND CHEMICAL EQUIPMENT REP	288	308	87	549	54D	63J
63J	88	QUARTERMASTER AND CHEMICAL EQUIPMENT REP	460	386	112	779	54D	63J
63J	89	QUARTERMASTER AND CHEMICAL EQUIPMENT REP	391	262	89	600	54D	63J
63N	85	M60A1/A3 TANK SYSTEM MECHANIC	105	791	.	.		63N
63N	86	M60A1/A3 TANK SYSTEM MECHANIC	105	704	.	.		63N
63N	87	M60A1/A3 TANK SYSTEM MECHANIC	87	505	.	.		63N
63N	88	M60A1/A3 TANK SYSTEM MECHANIC	51	269	.	.		63N
63S	86	HEAVY-WHEEL VEHICLE MECHANIC	92	1237	.	.		63S
63S	87	HEAVY-WHEEL VEHICLE MECHANIC	103	1335	.	.		63S
63S	88	HEAVY-WHEEL VEHICLE MECHANIC	112	1238	.	.		63S
63S	89	HEAVY-WHEEL VEHICLE MECHANIC	88	815	.	.		63S
63T	85	BRADLEY FIGHTING VEHICLE SYSTEM MECHANIC	60	1138	.	.		63T
63T	86	BRADLEY FIGHTING VEHICLE SYSTEM MECHANIC	53	839	.	.		63T
63T	87	BRADLEY FIGHTING VEHICLE SYSTEM MECHANIC	72	1104	.	.		63T
63T	88	BRADLEY FIGHTING VEHICLE SYSTEM MECHANIC	98	1500	.	.		63T
63T	89	BRADLEY FIGHTING VEHICLE SYSTEM MECHANIC	91	1317	.	.		63T

continued

Appendix D
(continued)

Sample Sizes for Army Specialties Used in the Analyses
By Selection Composite*

MOS	Year	Description	B	W	F	M	Prior	New
63W	85	WHEEL VEHICLE REPAIRER	276	916	66	1186		63W
63W	87	WHEEL VEHICLE REPAIRER	524	1458	124	1954		63W
63W	88	WHEEL VEHICLE REPAIRER	536	1371	117	1902		63W
6	89	WHEEL VEHICLE REPAIRER	414	860	77	1283		63W
67N	85	UTILITY HELICOPTER REPAIRER	44	852	.	.		67N
67N	86	UTILITY HELICOPTER REPAIRER	67	883	.	.		67N
67N	87	UTILITY HELICOPTER REPAIRER	62	860	.	.		67N
67N	88	UTILITY HELICOPTER REPAIRER	49	769	.	.		67N
67N	89	UTILITY HELICOPTER REPAIRER	45	670	.	.		67N
67T	87	TACTICAL TRANSPORT HELICOPTER REPAIRER	64	719	.	.		67T
67T	88	TACTICAL TRANSPORT HELICOPTER REPAIRER	68	759	.	.		67T
67T	89	TACTICAL TRANSPORT HELICOPTER REPAIRER	60	645	.	.		67T
67U	87	MEDIUM HELICOPTER REPAIRER	66	765	.	.		67U
67U	88	MEDIUM HELICOPTER REPAIRER	74	766	.	.		67U
67U	89	MEDIUM HELICOPTER REPAIRER	67	606	.	.		67U
67V	86	OBSERVATION/SCOUT HELICOPTER REPAIRER	53	642	.	.		67V
67V	87	OBSERVATION/SCOUT HELICOPTER REPAIRER	67	785	.	.		67V
67V	88	OBSERVATION/SCOUT HELICOPTER REPAIRER	73	881	.	.		67V
67V	89	OBSERVATION/SCOUT HELICOPTER REPAIRER	77	792	.	.		67V
67Y	87	AH-1 ATTACK HELICOPTER REPAIRER	44	650	.	.		67Y
67Y	88	AH-1 ATTACK HELICOPTER REPAIRER	40	617	.	.		67Y
68B	87	AIRCRAFT POWERPLANT REPAIRER	.	.	47	268		68B
68B	88	AIRCRAFT POWERPLANT REPAIRER	.	.	57	390		68B
68G	87	AIRCRAFT STRUCTURAL REPAIRER	42	472	.	.		68G
68G	88	AIRCRAFT STRUCTURAL REPAIRER	50	520	.	.		68G
		Operators and Food (OF) Composite						
13M	86	MULTIPLE LAUNCH ROCKET SYSTEM (MLRS) CRE	49	378			15D	13M
13M	87	MULTIPLE LAUNCH ROCKET SYSTEM (MLRS) CRE	51	555			15D	13M
13M	88	MULTIPLE LAUNCH ROCKET SYSTEM (MLRS) CRE	60	671			15D	13M
13M	89	MULTIPLE LAUNCH ROCKET SYSTEM (MLRS) CRE	79	607			15D	13M
13M	90	MULTIPLE LAUNCH ROCKET SYSTEM (MLRS) CRE	61	407			15D	13M
13N	88	LANCE CREUMEMBER	179	642			15D	13N
13N	89	LANCE CREWMEMBER	158	542			15D	13N
13N	90	LANCE CREUMEMBER	87	232			15D	13N
15D	85	MULTIPLE LAUNCH ROCKET SYSTEM (MLRS) CRE	66	620			15D	13M
15D	86	MULTIPLE LAUNCH ROCKET SYSTEM (MLRS) CRE	68	545			15D	13M
15D	87	MULTIPLE LAUNCH ROCKET SYSTEM (MLRS) CRE	80	518			15D	13M
15E	86	PERSHING MISSILE CREUMEMBER	135	670	.	.		15E
16D	85	HAWK MISSILE CREWMEMBER	45	235	56	230		16D
16D	86	HAWK MISSILE CREWMEMBER	94	345	66	391		16D
16D	87	HAWK MISSILE CREUMEMBER	94	342	69	383		16D
16D	88	HAWK MISSILE CREUMEMBER	91	361	52	414		16D
16D	89	HAWK MISSILE CREUMEMBER	79	274				16D
16E	85	HAWK FIRE CONTROL CREUMEMBER	70	332	.	.		16E
16E	86	HAWK FIRE CONTROL CREWMEMBER	122	472	89	524		16E
16E	87	HAWK FIRE CONTROL CREWMEMBER	111	391	63	455		16E
16E	88	HAWK FIRE CONTROL CREUMEMBER	63	191				16E
16P	85	CHAPARRAL CREUMEMBER	104	631				16P
16P	86	CHAPARRAL CREUMEMBER	170	944				16P
16P	87	CHAPARRAL CREWMEMBER	180	926				16P
16P	88	CHAPARRAL CREWMEMBER	146	787				16P

continued

Appendix D
(continued)

Sample Sizes for Army Specialties Used in the Analyses By Selection Composite*

MOS	Year	Description	B	W	F	M	Prior	New
16P	89	CHAPARRAL CREWMEMBER	106	506	.	.		16P
16R	85	VULCAN CREWMEMBER	47	399	.	.		16R
16R	86	VULCAN CREWMEMBER	80	635	.	.		16R
16R	87	VULCAN CREWMEMBER	103	707	.	.		16R
16R	88	VULCAN CREWMEMBER	119	704	.	.		16R
16R	89	VULCAN CREUMEMBER	71	317	.	.		16R
16S	85	MAN PORTABLE AIR DEFENSE SYSTEM CREUMEMB	399	714	.	.		16S
16S	86	MAN PORTABLE AIR DEFENSE SYSTEM CREUMEMB	558	926	.	.		16S
16S	87	MAN PORTABLE AIR DEFENSE SYSTEM CREUMEMB	528	830	.	.		16S
16S	88	MAN PORTABLE AIR DEFENSE SYSTEM CREUMEMB	575	889	.			16S
16S	89	MAN PORTABLE AIR DEFENSE SYSTEM CREUMEMB	530	733	.			16S
16S	90	MAN PORTABLE AIR DEFENSE SYSTEM CREUMEMB	333	380	.			16S
64C	85	MOTOR TRANSPORT OPERATOR	1863	4816	638	6267	64C	88M
64C	86	MOTOR TRANSPORT OPERATOR	2642	5915	921	7971	64C	88M
64C	87	MOTOR TRANSPORT OPERATOR	2739	5716	858	7909	64C	88M
88M	88	MOTOR TRANSPORT OPERATOR	2416	4969	810	6848	64C	88M
88M	89	MOTOR TRANSPORT OPERATOR	2057	3681	707	5235	64C	88M
88M	90	MOTOR TRANSPORT OPERATOR	1180	1816	366	2723	64C	88M
94B	85	FOOD SERVICE SPECIALIST	2236	2886	1040	4360		94B
94B	86	FOOO SERVICE SPECIALIST	2263	2480	1075	3904		948
94B	87	FOOD SERVICE SPECIALIST	2759	3080	1250	4871		94B
94B	88	FOOO SERVICE SPECIALIST	3268	3457	1267	5748		948
94B	89	FOOD SERVICE SPECIALIST	4008	3561	1378	6472		948
94F	85	HOSPITAL FOOO SERVICE SPECIALIST	45	114	67	99		94F
94F	86	HOSPITAL FOOD SERVICE SPECIALIST	70	150	114	118		94F
94F	87	HOSPITAL FOOD SERVICE SPECIALIST	95	183	138	154		94F
94F	88	HOSPITAL FOOD SERVICE SPECIALIST	95	161	135	133		94F
94F	89	HOSPITAL FOOD SERVICE SPECIALIST	57	86	74	77		94F

Surveillance and Communication (SC) Composite

MOS	Year	Description	B	W	F	M	Prior	New
31C	86	SINGLE CHANNEL RADIO OPERATOR	581	2732	266	3175	05B	31C
31C	87	SINGLE CHANNEL RADIO OPERATOR	718	3514	361	4051	05B	31C
31C	88	SINGLE CHANNEL RADIO OPERATOR	651	3126	333	3613	05B	31C
31C	89	SINGLE CHANNEL RADIO OPERATOR	881	3785	398	4464	05B	31C
72E	85	TACTICAL TELECOHHUNICATIONS CENTER OPERA	596	1276	358	1602	72F	72E
72E	86	TACTICAL TELECOMMUNICATIONS CENTER OPERA	605	1138	287	1542	72F	72E
72E	87	TACTICAL TELECOMMUNICATIONS CENTER OPERA	452	829	228	1118	72F	72E
72E	88	TACTICAL TELECOMMUNICATIONS CENTER OPERA	403	605	198	869	72F	72E
72E	89	TACTICAL TELECOMMUNICATIONS CENTER OPERA	332	434	169	640	72F	72E
72G	85	AUTOMATIC DATA TELECOMMUNICATIONS CENTER	188	509	247	466		72G
72G	86	AUTOMATIC DATA TELECOMMUNICATIONS CENTER	233	483	345	400		72G
72G	87	AUTOMATIC DATA TELECOMMUNICATIONS CENTER	328	580	459	499		72G
72G	88	AUTOMATIC DATA TELECOMMUNICATIONS CENTER	384	578	469	556		72G
72G	89	AUTOMATIC DATA TELECOMMUNICATIONS CENTER	284	341	290	380		72G
97G	86	COUNTER SIGNALS INTELLIGENCE SPECIALIST	.	.	47	109	05G	97G

* MOS is Military Occupational Specialty; Year is year tested; Prior and New refer to codes for the same specialty before and after the test data were collected.

Appendix E

Computational Formulas and Examples

The formulas used in each step of the analyses are provided in this appendix, along with sample results. Two Air Force classes were selected for use as samples: one a relatively large class using the Electronics (E) composite and the other a relatively small class using the Mechanical (M) composite. The notation used in this appendix is a blend of common statistical notation and variable names from the SAS programs used to process the data and compute the statistics of interest. Nearly all of the notation is explained in context.

A brief discussion of the ***unit of analysis*** may ***be*** helpful before proceeding to the detailed descriptions. Two levels of analyses are described:

- Individuals refer to individual recruits for whom both predictor (the ASVAB scores) and criterion (school grades or job performance) measures are available.
- A ***sample*** refers to a set of recruits for whom the ***exact same*** criterion measure is available. Each job necessarily involves a separate sample since each criterion measure applies to only one job.

In the case of the Army Skills Qualification Test (SQT) data, a new examination was created each year. Since the scores from different examinations for the same job were not carefully equated, it was necessary to treat the examinees taking different SQTs for the same job as separate samples. Thus, there were instances of multiple samples for the same job. There also were a few cases where the same individual was included in more than one sample, either because of repeated training courses or because the individual took more than one SQT. Such instances were relatively rare; consequently, the samples were treated as independent. In Step 2 below, the ***population*** is the 1980 Youth Population used for the ASVAB norms. The samples referred to were taken from subpopulations of the entire youth population, but it was not necessary to refer to these subpopulations in the text that follows.

In this appendix, the analyses are organized into the following steps:

- Estimate a criterion score for academic attritions;
- Adjust the criterion scales to a fixed estimated mean and standard deviation for the youth population as a whole;
- Compute regression equations for each sample and applicant group combination;
- Merge the regression equation statistics into a single file across the three Services;
- Compute the statistics of interest for each sample; and
- Aggregate across jobs and test statistical si cance.

The problem and approach for each step is described below, followed by the formulas, the SAS code, and sample results (as appropriate).

Step 1: Estimate a criterion score for academic attritions

Problem: Navy and Air Force results are based on training criteria. Recruits who did not complete training did not receive an appropriate final school grade (FSG). The use of the selection composite to predict whether a recruit will graduate is probably more important than the use to predict differences in final grades among the graduates. How can the dichotomous pass/fail outcome best be combined with the more continuous FSG outcome?

Approach: The modeled situation had the FSGs normally distributed for the combined sample of graduates and attritions; all students falling below a given score were academic attritions. Given the proportion passing, Pg, and the FSG mean and standard deviation for those passing, **MNg** and **SDg**, the mean score can be estimated as that score which those classed as academic attritions would have received, **MNa**; this mean can be assigned to all academic attrites.

Formula: If Pg is the percentage of recruits who graduate, then Z = **-NORMINV(Pg)** is the dividing point between attrites and graduates when the total distribution of FSG (including attrites) is standardized. Let Y = **f(Z)**, where **f()** is the normal density function so **f(t)** = **{1/sqrt(2*pi)}** * exp $(-t^2/2)$. For the remainder of this derivation, Y and Z are known values, computed as functions of the percentage of recruits who graduate, Pg.

In this total standardized metric, the mean score for the attrites is given by:

$$\mathbf{Ma} = \int_{-\infty}^{z} t\, f(t)\, dt \;\Big/\; \int_{-\infty}^{z} f(t)\, dt$$

Applying basic principles of calculus leads to Ma = **-Y/Pa**, where Pa = 1-Pg is the proportion of attrites.

Similarly, the mean score for graduates in this metric is given by:

$$\mathbf{Mg} = \int_{z}^{\infty} t\, f(t)\, dt \;\Big/\; \int_{z}^{\infty} f(t)\, dt = \mathbf{Y/Pg}.$$

In this same standardized metric, the variance of the scores for those passing is given by:

$$Vg = \int_{z}^{\infty} (t-Mg)^2\, f(t)\, dt \;\Big/\; \int_{z}^{\infty} f(t)\, dt.$$

A bit more calculus yields $Vg = 1 + Z\, Y/Pg - (Y/Pg)^2$.

Next, the translation between the observed FSG metric and the total standardized metric is derived. Let **MNg** and **SDg** be the observed mean and standard deviation for graduates. The translation is given by:

$$MNg = a*Mg + b \text{ and } SDg = a * sqrt(Vg).$$

So $a = SDg/sqrt\{1+ZY/Pg-(Y/Pg)^2\}$ and $b = MNg\text{-}a*Mg$.

Finally, **MNa**, the mean for attritions in the observed FSG metric, is given by:

$$MNa = a*Ma + b$$

which with a few substitutions and a little algebra becomes:

$$MNa = MNg - SDg*\{Y/(Pg*Pa)\}/sqrt\{1+ZY/Pg-(Y/Pg)^2\}.$$

SAS code:

```
Z=-PROBIT(PGRD);
Y=EXP(-.5*Z**2)/SQRT(2*3.14159);
A=(Y/(PGRD*(1-PGRD))) / SQRT(1 + Z*Y/PGRD -(Y/PGRD)**2);
ATTRMN = GRDMN - A*GRDSD;*** ASSIGNED SCORE FOR ATTRITES;
```

Sample results: The following shows actual values for two classes included in the analyses.

Class	ATTRN	GRDN	PGRD	Z	Y	A	GRDMN	GRDSD	ATTRMN
Samp1	195	1274	0.867257	-1.1135	0.214618	2.28919	90.4945	4.42308	80.3692
Samp2	3	291	0.989796	-2.3188	0.027126	2.77651	80.9725	7.06467	61.3574

Step 2. Adjust the criterion scale to a fixed estimated mean and standard deviation for the youth population as a whole

Problem: The approach to aggregation that was ultimately adopted involved the use of scale free statistics, so the scaling of the criterion variable within each sample does not matter to the tests for differences between applicant groups. For purposes of displaying composite prediction lines (averaged across different job samples) and for purposes of testing other aggregation methods, a common criterion scaling was desirable. Since the criterion samples were distinct and nonequivalent, it was not possible to compare the different criterion measures directly, but it was generally believed that the criterion measures for each course or job are on a scale that is influenced by the difficulty or complexity of the job. Getting a high grade in training for a complex and highly selective job is surely more difficult than getting a similar grade in a course open to nearly all recruits. Consequently, some adjustment for sample differences in examinee ability (and corresponding test difficulty) is desirable even though the important comparisons are not affected by differences in the criterion scale used with each sample.

Approach: The objective was to estimate an appropriate linear transformation of the criterion variable for each job/class sample so that the expected mean and variance for the entire (1980) youth population on the transformed scale would be the same for every sample. This would eliminate effects of differences in test difficulty and examinee abilities. The approach to identifying the appropriate transformation was to regress each criterion measure on the nine ASVAB subtests (with Paragraph Comprehension [PC] and

Word Knowledge [WK] combined into a single Verbal [VE] score) using the sample **data** and then to use the regression information to estimate the mean and variance for the youth population on the original criterion scale. The linear adjustment that would transform the youth population mean and standard deviation to the common target values was ident ed and used to adjust each criterion value. Initially, separate targets were selected for each Service to minimize the changes in the criterion score. Air Force school grades ranged from 0 to 100, with means averaging around 85 and standard deviations averaging around 5.0 across samples. The values 85 and 5 were chosen as the common mean and standard deviation targets for each Air Force sample. The same targets were also used for the Navy school grades. The Army SQT scores ranged from 0 to 100, but had an overall mean of about 75 and an average standard deviation of about 10, so 75 and 10 were used as the targets for the Army samples. In Step 4, the criterion measures were all rescaled to a mean of 0 and a standard deviation of 1 as the data for the different Services were combined. Note that no differentiation was made in Step 2 between the focal and reference applicant groups; the adjustments were based on each sample as a whole.

Formula: The multivariate range restriction correction attributed to **Lawley** (1943) in Lord and Novick (1968, p. 147) was used in estimating the population variance and mean on the existing criterion scale. The key formula for adjusting variances and covariances with this correction is:

$$\mathbf{C}_{pop} = \mathbf{C}_{samp} - \mathbf{V}' \; (\mathbf{P}^{-1}_{samp} - \mathbf{P}^{-1}_{samp} \; \mathbf{P}_{pop} \; \mathbf{P}^{-1}_{samp}) \; \mathbf{V}$$

where $\mathbf{C}_{pop}$ is the population covariance for a set of ***k*** criterion variables for which there was incidental selection due to correlation with explicit selection (predictor) variables (in this case there was only one criterion for each sample, so **k=1**); $\mathbf{C}_{samp}$ is the sample covariance for these variables; $\mathbf{P}_{samp}$ is the sample covariance matrix for the ***p*** explicit selection variables (in this case the nine ASVAB subtests); $\mathbf{P}_{pop}$ is the population covariance matrix for these same explicit selection variables (**from** the **NORC** study); and V is a ***pxk*** matrix of sample covariances for each combination of predictor and criterion variable. Note that if the implicit selection variables of interest are not affected by selection, then the covariance with each of the selection variables is zero; in this case the population and sample covariances are the same.

The above formula may also be rewritten as:

$$\mathbf{C}_{pop} = \mathbf{C}_{samp} \; - \; \mathrm{B} \; \mathbf{P}_{samp} \; \mathbf{B}' \; + \; \mathrm{B} \; \mathbf{P}_{pop} \; \mathbf{B}'$$

where $\mathrm{B} = \mathbf{V}' \; \mathbf{P}_{samp}^{-1}$ is a matrix of coefficients from the regression of the implicit selection variables (criteria) on the explicit selection variables (predictors). The correction thus amounts to subtracting out the covariance among the predicted values in the sample and replacing it with the covariance among the predicted values in the population. The residual of the covariances, uniqueness and error, is assumed to be independent of the selection and remains unchanged. The approach used in this adjustment makes no distributional assumptions. The underlying model assumes only that the regression is linear and that there is homogeneity of (prediction) error variances.

The full regression equation estimated from the sample is:

$$y_{pred} = B * \underline{x} + c_0$$

where y_{pred} is the predicted criterion value, B is the vector (matrix for multivariate criteria) of regression coefficients, $\underline{x}$ is a random vector of predictor (ASVAB) scores and c_0 is a constant (intercept) chosen so that the mean of the predicted values equals the observed sample criterion mean ($c_0 = My_{samp}$ - My_{pred}, where the My's are the means of the sample and predicted criterion values). Then substitute $\underline{Mx}_{pop}$, a vector of population ASVAB means, in the regression equation (for $\underline{x}$) to obtain an estimate of the population mean on the original criterion scale. Note that the equation for the population mean estimate can be written as:

$$My_{pop} = My_{samp} + B\ (\underline{Mx}_{pop} - \underline{Mx}_{samp})$$

where My_{pop} and My_{samp} are the mean criterion values for the population and sample, respectively, and $\underline{Mx}_{pop}$ and $\underline{Mx}_{samp}$ are vectors of predictor means for the population and sample.

Given estimates of the population mean and variance, My_{pop} and C_{pop}, on the original scale, then the adjustments are computed as:

$$a = TARGSD\ /\ Sqrt(C_{pop})$$

and $$b = TARGMN - a*My_{pop}$$

giving $$y_{adj} = a\ y_{orig} + b.$$

SAS code: The actual SAS (PROC MATRIX) code used to generate the estimates follows. Note that in this notation, POPCOVC and POPCRMN, are the target variance and mean for the adjusted scale, not the estimated values for the original scale.

```
CRITVAR=SAMPCOVS(ROW1+NPA:ROW1+NT1,NPA+1:NTOT);*ORDER=(NCxNC);
CRITSD=SQRT(DIAG(CRITVAR));*ORDER (NCXNC);
CSDI = INV(CRITSD);
ADJSMPV=SMPVAL*CSDI;  *PRED-CRIT COVS WITH STANDARDIZED CRIT;
SMPCRITV = POPCOVC*INV(IDC-ADJSMPV'*(SCOVPINV-SCOVPINV*POPCOVP
           * SCOVPINV) * ADJSMPV);
ADJCRSD = SQRT(VECDIAG(SMPCRITV))';
SAMPI = SAMPID(I,1);
OUTPUT ADJCRSD OUT=ADJCRSD ROWNAME=SAMPI COLNAME=CNAME2;
SMPPRMN = SAMPMNS(I,1:NPA);
ADJCRMN = POPCRMN + DIAG(ADJCRSD)*ADJSMPV'*SCOVPINV *
          (SMPPRMN - POPPRMN)';
```

Sample ***results***: The sample data that follow illustrate the computations. In general, each of the two samples shown has variances for the ASVAB subtests that are significantly smaller than the variances for the youth population. (The ASVAB subtest scores are all standardized to have a variance of 100 for the youth population.) Consequently, if the criterion is to be rescaled so that the youth population would have a standard deviation of 5.0 for the criterion, these selected samples would have somewhat smaller standard deviations (3.15 and 3.35). Also, the sample means on the relevant aptitude area composites are higher than the population mean. (The predictor composites are rescaled to have a mean of 100 and a standard deviation of 20.) If the criterion is scaled so that the youth population would have a mean of 85.0, then the target mean for these higher ability samples would be above 85.0 (89.2 and 86.7).

Population Covariance Matrix for the ASVAB Scores

GS	AR	VE	NO	CS	AS	MK	MC	EI
100	72	80	52	45	64	69	70	76
72	100	73	63	51	53	83	69	66
80	73	100	62	57	52	70	60	67
52	63	62	100	70	30	62	40	41
45	51	57	70	100	22	52	34	34
64	53	52	30	22	100	41	74	75
69	83	70	62	52	41	100	60	59
70	69	60	40	34	74	60	100	74
76	66	67	41	34	75	59	74	100

Sample Covariance Matrix for ASVAB Scores, Sample Class 1

SAMPCOVP	GS	AR	VE	NO	CS	AS	MK	MC	EI
GS	22.6	1.5	9.4	-1.8	-0.5	10.1	3.1	9.6	8.3
AR	1.5	20.3	3.7	6.8	6.1	4.4	10.2	8.8	-0.6
VE	9.4	3.7	14.3	1.4	3.6	3.7	2.9	4.8	3.8
NO	-1.8	6.8	1.4	36.1	22.0	-6.2	6.7	-1.9	-5.0
CS	-0.5	6.1	3.6	22.0	42.8	-5.4	5.7	-1.0	-3.9
AS	10.1	4.4	3.7	-6.2	-5.4	64.6	-4.7	28.9	29.5
MK	3.1	10.2	2.9	6.7	5.7	-4.7	27.2	5.7	-3.1
MC	9.6	8.8	4.8	-1.9	-1.0	28.9	5.7	46.3	17.1
EI	8.3	-0.6	3.8	-5.0	-3.9	29.5	-3.1	17.1	39.2

Covariance of Criterion with Predictors, Sample Class 1

V	GS	AR	VE	NO	CS	AS	MK	MC	EI	FSG
FSG	5.63	7.83	3.23	1.60	5.38	10.92	9.39	11.97	8 1	29.62

Inverse of the Sample ASVAB Covariance Matrix

SCOVPINV	GS	AR	VE	NO	CS	AS	MK	MC	EI
GS	0.068	0.008	-0.042	0.002	0.002	-0.004	-0.007	-0.006	-0.004
AR	0.008	0.072	-0.015	-0.007	-0.002	-0.007	-0.022	-0.010	0.007
VE	-0.042	-0.015	0.103	0.001	-0.008	0.003	0.001	0.001	-0.005
NO	0.002	-0.007	0.001	0.043	-0.020	0.001	-0.004	0.002	0.001
CS	0.002	-0.002	-0.008	-0.020	0.035	0.001	-0.001	0.000	0.000
AS	-0.004	-0.007	0.003	0.001	0.001	0.030	0.008	-0.012	-0.016
MK	-0.007	-0.022	0.001	-0.004	-0.001	0.008	0.050	-0.006	0.001
MC	-0.006	-0.010	0.001	0.002	0.000	-0.012	-0.006	0.035	-0.005
EI	-0.004	0.007	-0.005	0.001	0.000	-0.016	0.001	-0.005	0.042

The product SCOVPINV * POPCOV * SCOVPINV

SIPSI	GS	AR	VE	NO	CS	AS	MK	MC	EI
GS	0.14	0.07	-0.07	0.02	0.00	-0.01	0.01	-0.01	0.03
AR	0.07	0.18	0.03	0.02	-0.01	-0.02	0.01	-0.02	0.07
VE	-0.07	0.03	0.39	0.06	-0.01	0.01	0.05	-0.02	0.05
NO	0.02	0.02	0.06	0.09	-0.04	0.00	0.01	-0.00	0.02
CS	0.00	-0.01	-0.01	-0.04	0.06	0.00	0.00	0.00	-0.00
AS	-0.01	-0.02	0.01	0.00	0.00	0.03	0.01	-0.01	-0.02
MK	0.01	0.01	0.05	0.01	-0.00	0.01	0.08	-0.02	0.03
MC	-0.01	-0.02	-0.02	-0.00	0.00	-0.01	0.02	0.04	-0.01
EI	0.03	0.07	0.05	0.02	-0.00	-0.02	0.03	-0.01	0.08

Resulting values for both samples

SAMPID	TARGMN	TARGSD	SAMPMN	SAMPSD	ADJCOEF	ADJCONST
Sampl	89.2061	3.14861	89.1076	5.44208	0.578567	37.6514
Samp2	86.7119	3.34955	80.8362	7.23058	0.463247	49.2648

Statistics for the predictor (AASTD) and the original (FINALGRD) and adjusted (ADJGRD) criterion variables were as follows:

Predictor and Criterion Means (Before and After Adjustment) by AFS

AFS=Sampl

Variable	N	Mean	Standard Deviation	Minimum Value	Maximum Value	Skewness	Kurtosis
AASTD	1468	119.257	6.897	99.000	139.000	0.551	-0.442
FINALGRD	1468	89.108	5.442	76.000	99.000	-0.219	-0.899
ADJGRD	1468	89.204	3.149	81.621	94.926	-0.219	-0.899

AFS=Samp2

Variable	N	Mean	Standard Deviation	Minimum Value	Maximum Value	Skewness	Kurtosis
AASTD	293	109.229	10.897	87.000	140.000	0.635	-0.477
FINALGRD	293	80.836	7.231	61.000	98.000	-0.250	-0.318
ADJGRD	293	86.710	3.350	77.520	94.660	-0.250	-0.318

Note: AASTD is the aptitude composite rescaled to have a population mean of 100 with a standard deviation of 20, FINALGRD is the final school grade before rescaling the criterion, and ADJGRD is the final school grade adjusted to yield youth population means and standard deviation estimates at the targets.

For these samples, the predictor had some positive skewness due, primarily, to selection at the bottom end of the range. The criterion measures had some negative skewness, presumably due to a slight ceiling effect. The kurtosis was negative for both predictors and criterion due to some range restriction. These findings were typical of most of the training samples in the analyses. In the analyses that follow, the primary distributional assumption is that the distribution of the criterion conditional on the predictor measure was normal. Consequently, the skewness and kurtosis of the predictor measure were not an issue, but the conditional distribution of the criterion measure (i.e., of errors) was.

Step 3. Compute regression equations for each sample and applicant group combination

Problem: The next step was to estimate the relationship between criterion and predictor values separately for each sample and subgroup. As discussed in the report, a quadratic regression approach was used. In addition to generating an estimated criterion value at key points for each group, it was necessary to estimate the standard error of the estimated criterion values so that the significance of the differences could be determined.

Approach: An ordinary least squares (OLS) regression approach was used. The predictor variable was first rescaled so that the population mean would be zero in order to reduce the colinearity between the linear and quadriatic terms. Unfortunately, the sample means were mostly above the population mean so the two terms were substantially correlated in many samples. In the end (as seen in the examples), this correlation did not matter greatly since the primary concern was with the predicted values rather than with the regression coefficients.

SAS code: The SAS regression routine (PROC REG) estimates the variances and covariances among the parameter estimates (intercept and regression coefficients).

$$COV(b) = (X' X)^{-1} s^2$$

where X is the predictor data matrix (observations by variables) and s^2 is an estimate of the residual variance in the criterion after partialing out the variance predicted by the predictors.

Sample results: The data that follow show descriptive statistics and correlations, regression parameter estimates, and estimates of the covariance of these estimates for each of the two illustrative samples. The variable "PRDDEV" in the following output is the aptitude area composite rescaled by subtracting 100 and then dividing by 20.

Quadratic Regression Based on Air Force Training Data, by Race
Sampl, Reference Group (Whites)

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
PRDDEV	1218	0.991010	0.347203	1207.050000	0.450000	1.950000
PRDDEV2	1218	1.102551	0.760177	1342.907500	0.202500	3.802500
CRIT	1218	89.322227	3.181240	108794	81.621094	94.925781

Pearson Correlation Coefficients / N = 1218

Variable	PRDDEV	PRDDEV2	CRIT
PRDDEV	1.00000	0.98597	0.48359
PRDDEV2	0.98597	1.00000	0.48368
CRIT	0.48359	0.48368	1.00000

Samp2, Reference Group (Whites)

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
PRDDEV	1%	0.604523	0.553255	120.300000	-0.150000	2.000000
PRDDEV2	199	0.670000	0.851951	133.330000	0	4.000000
CRIT	199	87.323767	3.236619	17377	77.519531	94.660156

Pearson Correlation Coefficients / N = 199

Variable	PRDDEV	PRDDEV2	CRIT
PRDDEV	1.00000	0.93157	0.49423
PRDDEV2	0.93157	1.00000	0.48271
CRIT	0.49423	0.48271	1.00000

Sampl, Focal Group (Blacks)

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
PRDDEV	152	0.797368	0.244664	121.200000	0.500000	1.600000
PRDDEV2	152	0.695263	0-457304	105.680000	0.250000	2.560000
CRIT	152	88.289011	2.947338	13420	83.933594	94.925781

Pearson Correlation Coefficients / N = 152

Variable	PRDDEV	PRDDEV2	CRIT
PRDDEV	1.00000	0.98496	0.48681
PRDDEV2	0.98496	1.00000	0.48362
CRIT	0.48681	0.48362	1.00000

Samp2, Focal Group (Blacks)

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
PRDDEV	51	0.186275	0.352573	9.500000	-0.150000	1.250000
PRDDEV2	51	0.156569	0.307252	7.985000	0	1.562500
CRIT	51	85.486979	3.096653	4359.835938	78.910156	90.492188

Pearson Correlation Coefficients / N = 51

Variable	PRDDEV	PRDDEV2	CRIT
PRDDEV	1.00000	0.88833	0.34018
PRDDEV2	0.88833	1.00000	0.25845
CRIT	0.34018	0.25845	1.00000

Regression Parameter File Variables

OBS	COMPID	SAMPLE	TYPE	NAME	RMSE	INTERCEP	PRDDEV	PRDDEV2	SUBGRP	FREQ
1	E	Samp1	PARMS		2.78373	86.0016	2.2020	1.0325	W	1218
2	E	Samp1	COV	INTERCEP	2.78373	0.4832	-0.9372	0.4099	W	1218
3	E	Samp1	COV	PRDDEV	2.78373	-0.9372	1.8952	-0.8535	W	1218
4	E	Samp1	COV	PRDDEV2	2.78373	0.4099	-0.8535	0.3954	W	1218
5	E	Samp1	PARMS		2.59077	84.3023	4.2219	0.8922	B	152
6	E	Samp1	COV	INTERCEP	2.59077	4.7658	-10.7171	5.4997	B	152
7	E	Samp1	COV	PRDDEV	2.59077	-10.7171	24.8662	-13.1036	B	152
8	E	Samp1	COV	PRDDEV2	2.59077	5.4997	-13.1036	7.1177	B	152
9	M	Samp2	PARMS		2.82096	85.7023	1.9718	0.6410	W	199
10	M	Samp2	COV	INTERCEP	2.82096	0.1043	-0.1979	0.0826	W	199
11	M	Samp2	COV	PRDDEV	2.82096	-0.1979	0.9934	-0.6010	W	199
12	M	Samp2	COV	PRDDEV2	2.82096	0.0826	-0.6010	0.4189	W	199
13	M	Samp2	PARMS		2.95673	84.9563	4.6064	-2.0908	B	51
14	M	Samp2	COV	INTERCEP	2.95673	0.2216	-0.1779	-0.1086	B	51
15	M	Samp2	COV	PRDDEV	2.95673	-0.1779	6.6704	-6.7995	B	51
16	M	Samp2	COV	PRDDEV2	2.95673	-0.1086	-6.7995	8.7833	B	51

Step 4. Merge the regression equation statistics into a single file across the three Services

Problem: To this point, separate analyses were run for each Service to accommodate differences in editing requirements and the scaling of the variables. In order to merge results across Services, some rescaling of the variables, with corresponding adjustments to the parameter estimates, was required. In addition, the output from the regression program contained multiple lines (records) per sample. A consolidated file with one record per sample and subgroup was needed for aggregation.

Approach: The Air Force and Navy data were rescaled to have a criterion mean of zero and standard deviation of 1 in the youth population instead of 85 and 5. Army data were rescaled in a prior step. SAS code was created to retain the parameter estimates until all of the parameter covariance data were read in and then to output a single record per subgroup/sample combination.

SAS code:

```
SET IN1.AFPRMR(IN=INAF) IN2.NAVPRMR(IN=INNA);
BY COMPID SAMPLE SUBGRP;
RETAIN COO C01 C02 C11 C12 C22 A0 A1 A2 N 0;
IF FIRST.SUBGRP THEN DO;  A0=.;C00=.; C11=.; C22=.; END;
IF _TYPE_ EQ 'PARMS' THEN DO;
   A0=INTERCEP;  A1=PRDDEV; A2=PRDDEV2;N=_FREQ_;
END;
ELSE IF _NAME_ = 'INTERCEP' THEN DO;
     C00=INTERCEP;  C01=PRDDEV;  C02=PRDDEV2;
END;
ELSE IF _NAME_ EQ 'PRDDEV'  THEN DO;
   C11=PRDDEV;  C12=PRDDEV2;
END;
ELSE IF _NAME_ EQ 'PRDDEV2' THEN C22=PRDDEV2;
IF LAST.SUBGRP THEN DO;
   IF A0=. OR COO=. OR C11=. OR C22=. THEN ERROR 'MISSING';
        ***** STANDARDIZE CRITERION VARIABLE *****;
   A0=(A0-85)/5;  A1=A1/5;  A2=A2/5;  _RMSE_=_RMSE_/5;
   C00=C00/25; C11=C11/25; C22=C22/25;
   C01=C01/25; C02=C02/25; C12=C12/25;
   CRMN=(CRMN-85)/5; CRSD=CRSD/5;
   IF INAF THEN SRV='AF'; ELSE SRV='NA';
   OUTPUT;
END;
KEEP SRV COMPID SAMPLE SUBGRP N A O A1 A2  _RMSE_
     COO C11 C22 C01 C02 C12 PRMN PRSD CRMN CRSD;
```

Sample results: The output file for the two illustrative samples is shown below. Note that AO, A1, and **A2** are the intercept, linear, and quadratic coefficients respectively. **Cij** is the estimated covariance for the ith and jth parameter.

Fairness Analyses - Combined Race Results

Samp	/Subgrp	N	A0	A1	A2	RMSE	COO	C11	C22	C01	C02	C12
1	B	152	-0.140	0.844	0.178	0.518	0.191	0.995	0.285	-0.429	0.220	-0.524
1	W	1218	0.200	0.440	0.206	0.557	0.019	0.076	0.016	-0.037	0.016	-0.034
2	B	51	-0.009	0.921	-0.418	0.591	0.009	0.267	0.351	-0.007	-0.004	-0.272
2	W	199	0.140	0.394	0.128	0.564	0.004	0.040	0.017	-0.008	0.003	-0.024

Step 5. Compute the statistics of interest for each sample

Problem: At this stage, statistics indicating the differences between subgroups in the predictor-criterion relationships were computed. It was necessary to obtain estimates of both the size and the statistical significance of the differences for input into the routines that computed overall estimates of the size and statistical significance of the differences averaged across samples.

Appmach: The general approach to computing difference statistics involved several substeps: compute predicted criterion values at key points on the standardized predictor scale separately for the focal (black or female) and reference (white or male) groups and compute estimates of the standard errors of these predicted criterion values; compute differences in the predicted criterion values across applicant groups and compute estimates of the standard errors of these differences; and then compute a t value by dividing the estimated difference by its standard error. Because estimates of the standard errors that were pooled across applicant groups were not used, the degrees of freedom associated with this t value were not simple to compute. Based on the minimum sample size of 40 for each subgroup, it was appropriate to use a z approximation to the t value to summarize applicant group differences in the individual samples. Details and examples for each of these substeps follow.

Computing predicted values at key points. This was simply a matter of applying the regression parameters (intercept, linear, and quadratic coefficients) to the specified predictor values (population mean, mean plus and minus one-half standard deviation, and mean plus and minus one full standard deviation). For the first illustrative sample, the focal group parameter estimates were $\underline{b}$ = (A0,A1,A2) = (-.140, .844, .178). To obtain the estimated value at -.5 standard deviations, this vector was multiplied by $\underline{x}$ = (1,X,X^2) = (1, -.5, .25) to yield a predicted value of -.517. Similarly the reference group parameters, (.200, .440, .206), were multiplied by (1, -.5, .25) to yield a predicted value of .032.

Computing the standard error of the predicted values. Each predicted value was a linear composite of the estimated regression parameters. For the prediction at one-half standard deviation below the mean, for example, the vector product of the regression parameters, $\underline{b}$ = (AO, A1, A2), and $\underline{x}$ = (1, X, X^2) or (1, -.5, .25) was computed to get the predicted value. Since $\underline{x}$ is a fixed value, the variance of the predicted value is a function of the variance and covariance of the parameter estimates and was computed as $VAR(Y_{hat})$ = x' COV($\underline{b}$) x. (This follows the procedure outlined in the SAS 6.0 manual for computing standard errors for linear composites of regression parameter estimates.) For the focal group in the first illustrative sample at one-half standard deviation below the mean (AA=90), this computation was:

$$\underline{x}' * COV(\underline{b}) * \underline{x} = (\underline{x}'*COV(\underline{b})) * \underline{x} = VAR(Y_{hat})$$

$$(1,-.5,\ .25) * \begin{pmatrix} .191 & -.429, & .220 \\ -.429, & .995, & -.524 \\ .220 & -.524, & .285 \end{pmatrix} * \begin{pmatrix} 1.0 \\ -.5 \\ .25 \end{pmatrix} = (1,-.5,\ .25) * \begin{pmatrix} .461 \\ -1.058 \\ .553 \end{pmatrix} = 1.128$$

The standard error of the predicted focal group value is the square root of this variance or 1.062. For the reference group (White Males), the same computation led to a standard error of ,306 at one-half a standard deviation below the population mean on the predictor (AA=90).

SAS ***code:*** The SAS code used to compute the predicted values and their standard errors was:

```
DATA SUBGSTAT;
     SET IN.CMBPRMR2;
     RETAIN X80 80 X90 90 X100 100 X110 110 X120 120;
     ARRAY X  X80 X90 XlOO X110 X120;
     ARRAY YH YH80 YH90 YH100 YH110 YH120;
     ARRAY SE SE80 SE90 SE100 SE110 SE120;
     DO OVER X;
        Z=(X-100)/20;
```

```
        YH = A0 + A1*Z + A2*Z*Z;
        SE = SQRT(C00 + 2*C01*Z + (C11+2*C02)*Z**2
                      + 2*C12*Z**3 + C22*Z**4);
     END;
     SENS = YH120 - YH100;
     SE_SENS = SQRT(C11 + C22 + 2*C12);
     SE_RMSE = CRSD/SQRT(N);
     SE_A0=SQRT(C00); SE_A1=SQRT(C11); SE_A2=SQRT(C22);
     SEPM=PRSD/SQRT(N); SECM=CRSD/SQRT(N);
```

Computing Differences and Their Standard Errors. The next step was to compute sensitivity estimates for each group by taking the difference between the predicted value at one standard deviation above the population mean on the predictor (**AA**=120) and the predicted value at the predictor mean (**AA**=100). This can be expressed algebraically as:

$$\mathrm{SENS} = Y_{120}-Y_{100} = \underline{b}*\underline{x}_{120} - \underline{b}*\underline{x}_{100} = (A0,\ Al,\ A2)*(1,\ 1,\ 1)' - (A0,\ A1,\ A2)*(1,\ 0,\ 0)' = A1 + A2$$

so that the sensitivity measure was also a linear composite of the regression parameter estimates. The standard errors of the sensitivity measures were computed in the same way the standard errors for the predicted values were computed using:

$$\mathrm{VAR(sens)} = (\underline{x}_{120}-\underline{x}_{100})' * \mathrm{COV}(\underline{b}) * (\underline{x}_{120}-\underline{x}_{100}) = (0,1,1) * \mathrm{COV}(\underline{b}) * (0,1,1)' = C_{,,} + 2*C_{12} + C_{22}$$

The differences in predicted values for the focal and reference groups at each point (by subtraction) and the standard errors of these differences were also computed. Since the focal and reference groups were independent samples, the errors in estimating the regression parameters and hence the predicted values were uncorrelated so that the standard error of the differences was the square root of the sum of squares of the standard errors of the individual values. For example, for the first illustrative sample, the difference at one-half standard deviation below the population predictor mean and the standard error of this difference were computed as follows:

$$D = YF_{90} - YR_{90} = -.517 - .032 = -.549 \text{ and}$$
$$SE(D) = \mathrm{SQRT}(\ SE^2(YF_{90}) + SE^2(YR_{90})\ ) = \mathrm{SQRT}(\ 1.062^2 + .306^2) = 1.105$$

The difference in this example is in the direction of overprediction of black performance. Even with relatively large samples (1218 whites and 152 blacks) the standard error of this difference was quite large, so the obtained difference was clearly not statistically significant. The reason that the standard error was large for this difference (and the power to test the difference was so low) was that it is relatively removed from most of the data. The mean predictor values in standard deviation units were .80 for black males and .99 for white males with sample standard deviations of .25 and .35 respectively. The point in question, -.50 in standard deviation units, is more than four standard deviations ***below*** the sample means. At one standard deviation ***above*** the mean, the standard error of the difference in predicted values was only .069.

The SAS code to read the output from the prior step and compute the statistics of interest with their standard errors is as follows:

```
DATA COMBSTAT;
     SET SUBGSTAT;
     BY SRV COMPID SAMPLE SUBGRP;
     IF SRV EQ 'AF'  OR SRV EQ 'NA' THEN DO;
        PRMN=(PRMN-100)/20;  PRSD=PRSD/20;
      END;
     RETAIN WT80 WT90 WT100 WT110 WT120 N_REF N_FOC 0;
     RETAIN YF80 YF90 YF100 YF110 YF120 SF80 SF90 SF100 SF110 SF120 0;
     RETAIN YR80 YR90 YR100 YR110 YR120 SR80 SR90 SR100 SR110 SR120 0;
     RETAIN YD80 YD90 YD100 YD110 YD120 SD80 SD90 SD100 SD110 SD120 0;
     RETAIN SENF RMSE_F FO F1 F2 PRMF PRSF CRMF CRSF 0;
     RETAIN SF-SENS SF-RMSE SF-A0 SF-A1 SF-A2 SFPM SFCM 0;
     ARRAY WT   WT80--WT120 WTSN WT_ERR WO W1 W2 WTPM WTCM WTPS WTCS ;
     ARRAY STAT YH80--YH120 SENS _RMSE_ A0 A1 A2 PRMN CRMN PRSD CRSD ;
     ARRAY SE SE80--SE120 SE_SENS SE_RMSE SE_A0 SE_A1 SE_A2 SEPM SECM;
     ARRAY STAF YF80--YF120 SENF RMSE_F FO F1 F2 PRMF CRMF PRSF CRSF ;
     ARRAY SF SF80--SF120 SF-SENS SF-RMSE SF-A0 SF-A1 SF-A2 SFPM SFCM;
     ARRAY STAR YR80--YR120 SENR RMSE_R RO R1 R2 PRMR CRMR PRSR CRSR ;
     ARRAY SR SR80--SR120 SR_SENS SR_RMSE SR_A0 SR_A1 SR_A2 SRPM SRCM;
     ARRAY STAD YD80--YD120 SEND RMSE_D DO D1 D2 PRMD CRMD PRSD CRSD ;
     ARRAY SD SD80--SD120 SD_SENS SD_RMSE SD_A0 SD_A1 SD_A2 SDPM SDCM;
     IF FIRST.SAMPLE THEN DO;    **** COPY FOCAL GROUP VALUES TO RETAIN
VARS;
        DO OVER STAF; STAF=STAT; END;
        DO OVER SF;   SF=SE;     END;
        N_FOC=N;
      END;
      ELSE DO;   *** COMPUTE WEIGHTS FOR EACH VARIABLE AND SCALE THE;
                 *** VARIABLES SO DIFFERENCES ARE T SCORES;
         N_REF=N;
         IF N_FOC < 40 OR N_REF < 40 THEN DELETE;
         DO OVER STAR; STAR=STAT; STAD=STAF-STAR; END;
         DO OVER SR; SR=SE; SD=SQRT(SR**2+SF**2); END;
         DO OVER SD;   WT=1/SD;  END;  WTPS=WTPM; WTCS=WTCM;
         DO OVER STAR; STAR=STAR*WT; STAF=STAF*WT; STAD=STAD*WT; END;
         DO OVER SR; SR=(WT*SR)**2; SF=(WT*SF)**2; SD=(WT*SD)**2; END;
         OUTPUT;
      END;
```

After this step, the file contains the following values for the two illustrative samples.

COMPID	YF80	YR80	YD80	YF90	YR90	YD90	YF100	YR100	YD100
E	-0.805	-0.034	-0.772	-0.517	0.032	-0.549	-0.140	0.200	-0.340
M	-1.348	-0.126	-1.222	-0.574	-0.025	-0.549	-0.009	0.140	-0.149

COMPID	YF110	YR110	YD110	YF120	YR120	YD120
E	0.327	0.472	-0.145	0.883	0.847	0.036
M	0.347	0.370	-0.022	0.494	0.663	-0.169

COMPID	SF80	SR80	SD80	SF90	SR90	SD90	SF100	SR100	SD100
E	1.953	0.536	2.025	1.062	0.306	1.105	0.437	0.139	0.458
M	1.085	0.362	1.144	0.413	0.175	0.449	0.094	0.065	0.114

COMPID	SF110	SR110	SD110	SF120	SR120	SD120
E	0.086	0.038	0.094	0.066	0.022	0.069
M	0.142	0.054	0.152	0.245	0.058	0.252

COMPID	N_FOC	NREF	SENF	SENR	SEND	SF_SENS	SR_SENS	SD_SENS
E	152.0	1218.0	1.023	0.647	0.376	0.481	0.153	0.504
M	51.0	199.0	0.503	0.523	-0.019	0.272	0.092	0.287

COMPID	RMSE_F	RMSE_R	RMSE_D	SF_RMSE	SR_RMSE	SD_RMSE
E	0.518	0.557	-0.039	0.048	0.018	0.051
M	0.591	0.564	0.027	0.087	0.046	0.098

COMPID	PRMF	PRMR	PRMD	PRSF	PRSR	PRSD	CRMF	CRMR	CRMD
E	0.797	0.991	-0.194	0.245	0.347	-0.103	0.658	0.864	-0.207
M	0.186	0.605	-0.418	0.353	0.553	-0.201	0.097	0.465	-0.367

COMPID	CRSF	CRSR	CRSD	SFPM	SRPM	SDPM	SFCM	SRCM	SDCM
E	0.589	0.636	-0.047	0.397	0.199	0.444	0.403	0.187	0.444
M	0.619	0.647	-0.028	0.987	0.784	1.261	0.482	1.165	1.261

Step 6. Aggregate differences and standard errors across samples

Problem: The individual samples were too small to permit very powerful tests for subgroup differences. In addition, a meaningful summary of the overall impact of differences, across different jobs, was needed. Estimates of the statistical significance of aggregate difference estimates were also required.

Approach: The approach taken was to take a weighted average of the difference estimates from the individual samples. The weights used were the inverse of the standard errors of the differences. This amounted to taking a simple average of the z statistics (estimates divided by their standard error). Since the average was across literally hundreds of samples, the central limit theorem would indicate that the distribution of the average was extremely close to a normal distribution. (At this point, the z statistic from each sample could have been treated as a single observation, and a t test with degrees of freedom equal to the number of samples minus 1 could have been used to test whether the mean of these observations was significantly different from zero [again appealing to the central limit theorem]. However, the approach taken to computing the standard error of the average z value led to slightly greater precision.)

The weighted mean of the individual sample statistics was computed by summing the products of the individual sample statistics and their weights and then dividing this sum by the sum of the weights. The standard error of this weighted mean was computed as:

$$SE^2_{tot} = SUM(W^2_i * SE^2_i) \ / \ W^2_{tot}$$

where W_i is the weight given to sample i, SE_i is the standard error of the statistic in question for sample i, and W_{tot} is the sum of the weights across all samples. This is a very general formula that depends only on the assumption of independence for observations from the different samples.

As indicated above, the weights used were the inverse of the standard errors so that the mean difference from each sample was divided by its standard error, creating a scale-free Z statistic. Alternative weights were also explored ranging from unit weights to weights **defined** as the inverse of the square of the standard errors. The latter weights are optimal in the sense of minimizing the standard errors of the weighted means. The weights used in the main analyses were very nearly optimal and had the desirable property of removing any effects due to criterion scale differences.

After the weighted mean differences and their standard errors were computed, the hypothesis that the weighted mean was zero was tested against a two-tailed alternative. A Z approximation was used in this test since the exact degrees of freedom within, and hence across, samples was difficult to estimate. The degrees of freedom was quite large, as several hundred samples were included, so that a **normal** approximation was quite satisfactory.

SAS code:

```
 PROC MEANS MAXDEC=3  DATA=COMBSTAT;
 VAR N_FOC N_REF
        WT80--WT120 WTSN WT_ERR WO W1 W2 WTPM WTPS WTCM WTCS
        YD80--YD120 SEND RMSE_D DO D1 D2 PRMD PRSD CRMD CRSD
        SD80--SD120 SD_SENS SD_RMSE SD_A0 SD_A1 SD_A2 SDPM SDCM
        YF80--YF120 SENF RMSE_F FO F1 F2 PRMF PRSF CRMF CRSF
        SF80--SF120 SF-SENS SF_RMSE SF-A0 SF-A1 SF-A2 SFPM SFCM
        YR80--YR120 SENR RMSE_R RO R1 R2 PRMR PRSR CRMR CRSR
        SR80--SR120 SR-SENS SR_RMSE SR_A0 SR_A1 SR_A2 SRPM SRCM;
 OUTPUT OUT=OUTMEAN0
 MEAN=N_FOC N_REF
 WT80 WT90 WT100 WT110 WT120 WTSN WT_ERR WO W1 W2 WTPM WTPS WTCM WTCS
 YD80 YD90 YD100 YD110 YD120 SEND RMSE_D DO D1 D2 PRMD PRSD CRMD CRSD
 SD80 SD90 SD100 SD110 SD120 SD_SENS SD_RMSE SD_A0 SD_A1 SD_A2 SDPM SDCM
 YF80 YF90 YF100 YF110 YF120 SENF RMSE_F FO F1 F2 PRMF PRSF CRMF CRSF
 SF80 SF90 SF100 SF110 SF120 SF-SENS SF-RMSE SF-A0 SF-A1 SF-A2 SFPM SFCM
 YR80 YR90 YR100 YR110 YR120 SENR RMSE_R RO R1 R2 PRMR PRSR CRMR CRSR
 SR80 SR90 SR100 SR110 SR120 SR-SENS SR-RMSE SR_A0 SR_A1 SR_A2 SRPM SRCM
   N=NSAMPS;
```

Note: Means were saved rather than sums. For estimating overall means, the differences (a factor of one over the number of samples) canceled out when the mean of the weight times statistic values was divided by the mean of the weight values. In computing standard errors, it was necessary to modify the formula slightly to accommodate the use of means.

```
DATA RESULTS;
  SET OUTMEAN0;
  ARRAY WT   WT80--WT120 WTSN WT_ERR WO W1 W2 WTPM WTPS WTCM WTCS ;
  ARRAY STAF' YF80--YF120 SENF RMSE_F FO F1 F2 PRMF PRSF CRMF CRSF ;
  ARRAY SF SF80--SF120 SF-SENS SF-RMSE SF-A0 SF-A1 SF-A2 SFPM SFCM;
  ARRAY STAR YR80--YR120 SENR RMSE_R RO R1 R2 PRMR PRSR CRMR CRSR ;
  ARRAY SR SR80--SR120 SR-SENS SR-RMSE SR_A0 SR_A1 SR_A2 SRPM SRCM;
  ARRAY STAD YD80--YD120 SEND RMSE_D DO D1 D2 PRMD PRSD CRMD CRSD ;
  ARRAY SD SD80--SD120 SD_SENS SD_RMSE SD_A0 SD_A1 SD_A2 SDPM SDCM;
  DO OVER STAD; STAR=STAR/WT; STAF=STAF/WT; STAD=STAD/WT; END;
  DO OVER SD;
     SR=SQRT(SR/(NSAMPS*WT**2)); SF=SQRT(SF/(NSAMPS*WT**2));
     SD=SQRT(SR**2+SF**2);
  END;
```

```
PROC PRINT DATA=RESULTS;
TITLE3 'MEAN  STANDARDIZED CRITERION LEVELS FOR KEY PREDICTOR LEVELS';
    TITLE4 'OVERALL  AND BY COMPOSITE';
    ID SRV COMPID;
    VAR YF80 YR80 YD80  YF90 YR90 YD90
         YF100 YR100 YD100  YF110 YR110 YD110
         YF120 YR120 YD120 NSAMPS;
    FORMAT YF80--YF120 YR80--YR120  YD80--YD120 6.3;

PROC PRINT DATA=RESULTS;
TITLE3 'STANDARD  ERRORS FOR PREDICTED CRITERION LEVELS';
    TITLE4 'OVERALL  AND BY COMPOSITE';
    ID SRV COMPID;
    VAR SF80 SR80 SD80  SF90 SR90 SD90
         SF100 SR100 SD100  SF110 SR110 SD110
         SF120 SR120 SD120 NSAMPS;
    FORMAT SF80--SF120 SR80--SR120  SD80--SD120 6.3;

PROC PRINT DATA=RESULTS;
TITLE3 'SENSITIVITY  AND PREDICTION ERROR LEVELS';
    TITLE4 'OVERALL  AND BY COMPOSITE';
    ID SRV COMPID;
    VAR NSAMPS N_FOC N_REF SENF SENR SEND SF-SENS SR_SENS SD_SENS
         RMSE_F RMSE_R RMSE_D SF_RMSE SR_RMSE SD_RMSE;
    FORMAT SENF SENR SEND SF-SENS SR_SENS SD_SENS 6.3
             RMSE_F RMSE_R RMSE_D SFRMSE SR_RMSE SD_RMSE 6.3;

PROC PRINT DATA=RESULTS;
    TITLE3 'PREDICTOR  AND CRITERION MEANS';
    TITLE4 'OVERALL  AND BY COMPOSITE';
    ID SRV COMPID;
    VAR PRMF PRMR PRMD PRSF PRSR PRSD CRMF CRMR CRMD CRSF CRSR CRSD
         SFPM SRPM SDPM SFCM SRCM SDCM;
FORMAT PRMF PRMR PRMD PRSF PRSR PRSD CRMF CRMR CRMD CRSF CRSR CRSD
         SFPM SRPM SDPM SFCM SRCM SDCM  6.3;

PROC PRINT DATA=RESULTS;
    TITLE3 'PREDICTION  PARAMETER ESTIMATES';
    TITLE4 'OVERALL  AND BY COMPOSITE';
    ID SRV COMPID;
    VAR FO SF-A0 RO SR_A0 DO SD_A0 F1 SF-A1 R1 SR_A1 D1 SD_A1
         F2 SF_A2 R2 SR_A2 D2 SD_A2;
FORMAT FO SF_A0 RO SR_A0 DO SD_A0 F1 SF-A1 R1 SR_A1 D1 SD_A1
         F2 SF-A2 R2 SR_A2 D2 SD_A2  6.3;
```

Sample results:

Fairness Results by Race
Model: Quadratic Wts: T-Vals
Reference Group: Whites Focal Group: Blacks
Min N: 40

Variable	N	Mean	Std Dev	Minimum	Maximum
N_FOC	2	101.500	71.418	51.000	152.000
N_REF	2	708.500	720.542	199.000	1218.000
WT80	2	0.684	0.269	0.494	0.874
WT90	2	1.567	0.936	0.905	2.229
WT100	2	5.471	4.651	2.182	8.759
WT110	2	8.615	2.883	6.577	10.654
WT120	2	9.198	7.395	3.969	14.427
WTSN	2	2.731	1.058	1.983	3.479
WT_ERR	2	14.867	6.612	10.192	19.543
W0	2	5.471	4.651	2.182	8.759
W1	2	1.386	0.594	0.967	1.806
W2	2	1.736	0.124	1.648	1.824
WTPM	2	1.523	1.032	0.793	2.252
WTPS	2	1.523	1.032	0.793	2.252
WTCM	2	22.523	7.810	17.000	28.045
WTCS	2	22.523	7.810	17.000	28.045
YD80	2	-0.725	0.486	-1.069	-0.381
YD90	2	-0.861	0.515	-1.225	-0.497
YD100	2	-1.024	0.400	-1.307	-0.742
YD110	2	-0.845	0.988	-1.544	-0.147
YD120	2	-0.075	0.841	-0.669	0.520
SEND	2	0.339	0.575	-0.068	0.745
RMSE_D	2	-0.239	0.729	-0.754	0.277
D0	2	-1.024	0.400	-1.307	-0.742
D1	2	0.671	0.397	0.390	0.952
D2	2	-0.476	0.601	-0.901	-0.051
PRMD	2	-0.384	0.074	-0.436	-0.332
PRSD	2	-0.195	0.051	-0.231	-0.159
CRMD	2	-6.908	4.801	-10.303	-3.513
CRSD	2	-0.790	0.007	-0.795	-0.785
SD80	2	1.000	0.000	1.000	1.000
SD90	2	1.000	0.000	1.000	1.000
SD100	2	1.000	0.000	1.000	1.000
SD110	2	1.000	0.000	1.000	1.000
SD120	2	1.000	0.000	1.000	1.000
SD_SENS	2	1.000	0.000	1.000	1.000
SD_RMSE	2	1.000	0.000	1.000	1.000
SD_A0	2	1.000	0.000	1.000	1.000
SD_A1	2	1.000	0.000	1.000	1.000
SD_A2	2	1.000	0.000	1.000	1.000
SDPM	2	1.000	0.000	1.000	1.000
SDCM	2	1.000	0.000	1.000	1.000
YF80	2	-0.788	0.552	-1.179	-0.398
YF90	2	-0.874	0.574	-1.280	-0.468
YF100	2	-0.191	0.161	-0.305	-0.077
YF110	2	2.886	0.850	2.284	3.487
YF120	2	7.353	7.624	1.962	12.743
SENF	2	1.889	0.196	1.751	2.028
RMSE_F	2	8.077	2.899	6.027	10.126

continued

Fairness Results by Race (continued)
Model: Quadratic Wts: T-Vals
Reference Group: Whites Focal Group: Blacks
Min N: 40

Variable	N	Mean	Std Dev	Minimum	Maximum
F0	2	-0.191	0.161	-0.305	-0.077
F1	2	1.240	0.600	0.816	1.664
F2	2	-0.182	0.718	-0.689	0.325
PRMF	2	0.972	1.165	0.148	1.796
PRSF	2	0.415	0.192	0.280	0.551
CRMF	2	6.957	5.976	2.731	11.183
CRSF	2	13.695	5.196	10.021	17.369
SF80	2	0.915	0.021	0.900	0.930
SF90	2	0.885	0.054	0.847	0.923
SF100	2	0.794	0.161	0.680	0.908
SF110	2	0.856	0.028	0.836	0.875
SF120	2	0.923	0.035	0.898	0.947
SF-SENS	2	0.903	0.007	0.898	0.908
SF_RMSE	2	0.827	0.065	0.781	0.873
SF-A0	2	0.794	0.161	0.680	0.908
SF-A1	2	0.900	0.042	0.870	0.929
SF-A2	2	0.951	0.005	0.947	0.954
SFPM	2	0.706	0.132	0.613	0.799
SFCM	2	0.484	0.478	0.146	0.823
YR80	2	-0.063	0.066	-0.110	-0.017
YR90	2	-0.013	0.059	-0.055	0.029
YR100	2	0.834	0.561	0.437	1.230
YR110	2	3.731	1.838	2.431	5.030
YR120	2	7.427	6.783	2.631	12.223
SENR	2	1.550	0.379	1.283	1.818
RMSE_R	2	8.315	3.627	5.750	10.880
RO	2	0.834	0.561	0.437	1.230
R1	2	0.569	0.203	0.426	0.712
R2	2	0.294	0.117	0.211	0.377
PRMR	2	1.356	1.239	0.479	2.232
PRSR	2	0.610	0.243	0.439	0.782
CRMR	2	13.865	1.175	13.034	14.695
CRSR	2	14.485	5.189	10.816	18.154
SR80	2	0.085	0.021	0.070	0.100
SR90	2	0.115	0.054	0.077	0.153
SR100	2	0.206	0.161	0.092	0.320
SR110	2	0.144	0.028	0.125	0.164
SR120	2	0.077	0.035	0.053	0.102
SR_SENS	2	0.097	0.007	0.092	0.102
SR_RMSE	2	0.173	0.065	0.127	0.219
SR_A0	2	0.206	0.161	0.092	0.320
SR_A1	2	0.100	0.042	0.071	0.130
SR_A2	2	0.049	0.005	0.046	0.053
SRPM	2	0.294	0.132	0.201	0.387
SRCM	2	0.516	0.478	0.177	0.854

Mean Standardized Criterion Levels for Focal (YF') and Reference (YR) Groups and Subgroup Differences (YD) at Key Predictor Levels, Overall and by Composite

COMPID	YF80	YR80	YD80	YF90	YR90	YD90	YF100	YR100	YD100
Total	-1.152	-0.092	-1.060	-0.558	-0.008	-0.549	-0.035	0.152	-0.187
E	-0.805	-0.034	-0.772	-0.517	0.032	-0.549	-0.140	0.200	-0.340
M	-1.348	-0.126	-1.222	-0.574	-0.025	-0.549	-0.009	0.140	-0.149

COMPID	YF110	YR110	YD110	YF120	YR120	YD120	NSMPS
Total	0.335	0.433	-0.098	0.799	0.807	-0.008	2
E	0.327	0.472	-0.145	0.883	0.847	0.036	1
M	0.347	0.370	-0.022	0.494	0.663	-0.169	1

Standard Errors for Focal (SF'), Reference (SR), and Difference (YD) Statistics

COMPID	SF80	SR80	SD80	SF90	SR90	SD90	SF100	SR100	SD100
Total	0.989	0.302	1.034	0.424	0.153	0.451	0.115	0.059	0.129
E	1.953	0.536	2.025	1.062	0.306	1.105	0.437	0.139	0.458
M	1.085	0.362	1.144	0.413	0.175	0.449	0.094	0.065	0.114

COMPID	SF110	SR110	SD110	SF120	SR120	SD120	NSMPS
Total	0.076	0.031	0.082	0.074	0.021	0.077	2
E	0.086	0.038	0.094	0.066	0.022	0.069	1
M	0.142	0.054	0.152	0.245	0.058	0.252	1

Sensitivity and Prediction Error Levels, Overall and by Composite

COMPID	NSAMPS	N FOC	N REF	SENF	SENR	SEND	SF SENS	SR SENS	SD SENS
Total	2	101.5	708.5	0.692	0.568	0.124	0.246	0.081	0.259
E	1	152.0	1218.0	1.023	0.647	0.376	0.481	0.153	0.504
M	1	51.0	199.0	0.503	0.523	-0.019	0.272	0.092	0.287

COMPID	RMSE F	RMSE R	RMSE D	SF RMSE	SR RMSE	SD RMSE
Total	0.543	0.559	-0.016	0.043	0.020	0.048
E	0.518	0.557	-0.039	0.048	0.018	0.051
M	0.591	0.564	0.027	0.087	0.046	0.098

Predictor and Criterion Means, Overall and by Composite

COMPID	PRMF	PRMR	PRMD	PRSF	PRSR	PRSD	CRMF	CRMR	CRMD
Total	0.638	0.890	-0.252	0.273	0.401	-0.128	0.309	0.616	-0.307
E	0.797	0.991	-0.194	0.245	0.347	-0.103	0.658	0.864	-0.207
M	0.186	0.605	-0.418	0.353	0.553	-0.201	0.097	0.465	-0.367

COMPID	CRSF	CRSR	CRSD	SFPM	SRPM	SDPM	SFCM	SRCM	SDCM
Total	0.608	0.643	-0.035	0.390	0.252	0.464	0.323	0.333	0.464
E	0.589	0.636	-0.047	0.397	0.199	0.444	0.403	0.187	0.444
M	0.619	0.647	-0.028	0.987	0.784	1.261	0.482	1.165	1.261

Prediction Parameter Estimates, Overall and by Composite

Intercept

	Focal		Reference		Difference	
COMPID	Parm	SE	Parm	SE	Parm	SE
Total	-0.035	0.115	0.152	0.059	-0.187	0.129
E	-0.140	0.437	0.200	0.139	-0.340	0.458
M	-0.009	0.094	0.140	0.065	-0.149	0.114

Linear Coefficient

	Focal		Reference		Difference	
COMPID	Parm	SE	Parm	SE	Parm	SE
Total	0.894	0.484	0.410	0.161	0.484	0.510
E	0.884	0.997	0.440	0.275	0.404	1.035
M	0.921	0.517	0.394	0.199	0.527	0.554

Quadratic Coefficient

	Focal		Reference		Difference	
COMPID	Parm	SE	Parm	SE	Parm	SE
Total	-0.105	0.397	0.169	0.090	-0.274	0.407
E	0.178	0.534	0.206	0.126	-0.028	0.548
M	-0.418	0.593	0.128	0.129	-0.546	0.607