Establishing Gender-Neutral Physical Standards for Ground Combat Occupations

Volume 2. A Review of the Military Services' Methods

Chaitra M. Hardison, Susan D. Hosek, Anna Rosefsky Saavedra

RAND CORPORATION

For more information on this publication, visit www.rand.org/t/RR1340z2

Published by the RAND Corporation, Santa Monica, Calif.

Cover: U.S. Army image from Reuters.

Preface

On January 24, 2013, the Secretary of Defense and Chairman of the Joint Chiefs of Staff announced rescission of the 1994 Direct Ground Combat Definition and Assignment Rule, which restricted assignments of women to occupational specialties or positions in or collocated with direct ground combat units. After the announced decision to eliminate the rule, the fiscal year 2014 National Defense Authorization Act (NDAA) directed establishment of gender-neutral and valid physical standards prior to opening any of the restricted occupations to women and gave the services until no later than September 2015 to demonstrate that such standards were in place.

This report provides the results of a review to assess whether the work undertaken by the services would satisfy the NDAA requirements. To accomplish this, the research proceeded in two phases. The first phase involved describing best-practice methodologies for establishing gender-neutral standards for physically demanding jobs, tailored to address the needs of the military. The work in the first phase was completed in 2013, and the report (Volume 1) was shared with the services. The second phase of the work, described in this report (Volume 2), involved a review and evaluation of the methods used by the military services to meet the requirement. The review began in 2013, as the services planned their efforts, and ended in April 2015, about six months prior to the 2015 deadline and before the services had completed the final stages of their work.

In December 2015, the Office of the Secretary of Defense released a preliminary version of this report on its website when the Secretary of Defense announced the decision to open previously closed combat

occupations to women. This final version of the report incorporates minor changes resulting from RAND's quality assurance process and has been copyedited and proofread for clarity and ease of reading.

This research was sponsored by the Office of the Under Secretary of Defense for Personnel and Readiness. It was conducted within the Forces and Resources Policy Center of the RAND National Defense Research Institute, a federally funded research and development center sponsored by the Office of the Secretary of Defense, the Joint Staff, the Unified Combatant Commands, the Navy, the U.S. Marine Corps, the defense agencies, and the defense Intelligence Community.

For more information on the RAND Forces and Resources Policy Center, see www.rand.org/nsrd/ndri/centers/frp or contact the director (contact information is provided on the webpage).

Contents

Preface iii
Figures ix
Tables xi
Summary xiii
Acknowledgments xxxv
Abbreviations xxxvii

CHAPTER ONE
Introduction 1
Setting the Stage 5
The Importance of Establishing Valid Physical Standards in Light of the Policy Change 6
Organization of This Report 7

CHAPTER TWO
Recommended Process for Establishing Physical Standards 9
Stage 1. Identify Physical Demands (Job Analysis) 11
Stage 2. Identify Potential Screening Tests 12
Stage 3. Validate and Select Tests 13
Stage 4. Establish Minimum Scores 14
Stage 5. Implement Screening 15
Stage 6. Confirm Tests Are Working as Intended 16
Summary 17

CHAPTER THREE
The Analytic Approach for Evaluating the Services' Efforts 19

CHAPTER FOUR
Army Combat Arms 25
Occupational Assignment and Screening in the Army 26
Army's Process for Establishing Standards for Combat Arms 28
Our Evaluation 40

CHAPTER FIVE
Army Special Operations Forces 43
Occupational Assignment and Screening in USASOC 45
Army's Process for Establishing Standards for Special Operations Forces 50
Our Evaluation 52

CHAPTER SIX
Marine Corps Combat Arms 57
Occupational Assignment and Screening in the Marine Corps 57
Marine Corps' Process for Establishing Standards for Combat Arms 60
Our Evaluation 90

CHAPTER SEVEN
Marine Corps Special Operations Forces 93
Occupational Assignment and Screening 93
MARSOC's Process for Establishing Standards for Special Operations Forces 96
Our Evaluation 100

CHAPTER EIGHT
Navy Special Operations Forces 103
Occupational Assignment and Screening 103
Navy's Process for Validating Existing Standards for Special Operations Forces 110
Our Evaluation 121

CHAPTER NINE
Air Force Battlefield Airmen 125
Occupational Assignment and Screening in the Air Force 126
Air Force Process for Establishing Standards for Battlefield Airmen 131
Our Evaluation 138

CHAPTER TEN
Overarching Observations, Findings, and Recommendations 141
Comparing the Services' Efforts 141
Remaining Stages of the Standards Setting Process 147
Final Thoughts 151

APPENDIXES
A. Terminology Used in Setting Physical Standards 153
B. Physically Demanding Occupations Open to Women Before 2016 165

References 179

Figures

S.1. Six Stages in Developing Physical Standards xvi
2.1. Six Stages in Developing Physical Standards 11
3.1. Physical Standards Development Process 21
4.1. Eligibility and Training Path for Enlisted Personnel in Army Closed Occupations 27
4.2. Illustration of a Hypothetical Selection Test Biased in Favor of Male Applicants 39
5.1. Eligibility and Training Path for Army Special Forces 46
6.1. Eligibility and Training Path for Marine Corps Combat Arms Branches 58
6.2. Study Population Used to Calculate Cutoff Scores for Infantry Qualifying Test Proposed in NHRC-TECOM Report 71
6.3. Hypothesized Relationship Between Individual and Unit Attributes in the Ground Combat Element Integrated Task Force Study 81
7.1. Eligibility and Training Path for Marine Corps Special Operations Forces 94
8.1. Eligibility and Training Requirements for Navy SEALs 105
8.2. Eligibility and Training Requirements for Navy SWCCs 107
9.1. Eligibility and Training Requirements for Enlisted Battlefield Airmen 130
9.2. Incrementally Higher Minimums Account for Improvement Gained from Training 136
B.1. Eligibility and Training Path for Navy EOD Technicians 170
B.2. Eligibility and Training Path for Navy Divers 171
B.3. Eligibility and Training Path for Navy AIRRs 172

Tables

S.1. Summary of Key Features of the Service Approaches xix
4.1. Physical Tasks for Infantry (11B) 28
4.2. Initial Task List Used in the Army Simulation Observation Study 30
6.1. Marine Corps Ground Combat Enlisted Occupations with Physically Demanding Tasks and Closed to Women 64
6.2. Proxy Tasks for Physically Demanding Tasks in Marine Corps Occupations Closed to Women 67
6.3. Mission Events by Occupation for GCEITF Study 84
6.4. Performance Measures for Ground Combat Element Integrated Task Force Study 86
8.1. Test Scores of SEAL and SWCC Graduates and Nongraduates 113
9.1. Physical Ability Stamina Test Minimums for Enlisted Jobs 128
10.1. Summary of Key Features of the Service Approaches 142
B.1. Minimum PST Scores 169
B.2. Female AIRR Recruiting Goal 175

Summary

Although the role of women in the military has been gradually expanding since World War II, women have been banned from serving in specialties and assignments that involve direct combat on the ground during much of this period. However, on January 24, 2013, Secretary of Defense Leon Panetta and Chairman of the Joint Chiefs of Staff GEN Martin Dempsey sent a memo to the military services rescinding the 1994 Direct Ground Combat Definition and Assignment Rule (herein called the "ground combat exclusion policy") and announcing the intention to "integrate women into occupational fields to the maximum extent possible" as of January 2016 (Panetta and Dempsey, 2013).

As the military opened new positions to women, particularly positions with physically demanding tasks, the services needed a more systematic way to determine who would be qualified to fill these positions. Section 543 of the National Defense Authorization Act (NDAA) for 1994 mandated gender-neutral occupational standards to qualify individuals for any military occupation open to men and women and gender-neutral "specific physical requirements" for open occupations in which performance depends on "muscular strength and endurance and cardiovascular capacity" (Pub. L. 103-160, 1993). The fiscal year (FY) 2015 NDAA required that the "gender-neutral occupational standards being developed by the secretaries of the military departments (1) accurately predict performance of actual, regular, and recurring duties of a military occupation; and (2) are applied equitably to measure individual capabilities" (Pub. L. 113-291, 2014). These gender-neutral standards were to be developed, reviewed, and validated no later than

September 2015, as specified in Section 524 of the FY 2014 NDAA (Pub. L. 113-66, 2013).

Mindful of these responsibilities, the Office of the Under Secretary of Defense for Personnel and Readiness asked RAND researchers to help it understand how to evaluate job-specific physical requirements and establish gender-neutral standards for physically demanding jobs. Our study addressed two research objectives. The first was to describe best-practice methodologies for establishing gender-neutral standards for physically demanding jobs, tailored to address the needs of the military. The second objective of the study was to review and evaluate methodologies being used by the military services to set gender-neutral standards. This report provides the results of work conducted toward the second research objective, using the best-practice methodology described in Volume 1 of this study as a framework. The review began in 2013, as the services planned their efforts, and ended in April 2015, about six months prior to the 2015 deadline and before the services had completed the final stages of their work. Volume 1 provides detailed explanations of terms and concepts used in this document.

In this report, we use the term *standards* or *physical standards* to refer to occupation-specific criteria that applicants must meet to enter or remain in a particular career field or specialty. More specifically, we are concerned with standards used to make selection decisions—that is, decisions made that could exclude people from entering or continuing in a job. With respect to the term *gender-neutral standards,* we use a simple and straightforward definition. If the minimum passing score is the same for women as it is for men, then it is *gender-neutral*. Nevertheless, according to the NDAA, the intention is not merely that a physical selection standard be the same for both genders, but that the scores on screening tests also be *valid* (i.e., useful in predicting success in critical physical requirements of the job) and applied equitably.

Exploring the effects on subgroups, such as gender and race/ethnicity, is one important way to ensure validity and equitable treatment in selection practices. In particular, test validity should not differ among relevant subgroups, and test scores should be unbiased (i.e., two people who receive the same test score should have the same likelihood of success on the job, regardless of subgroup). It is worth noting that

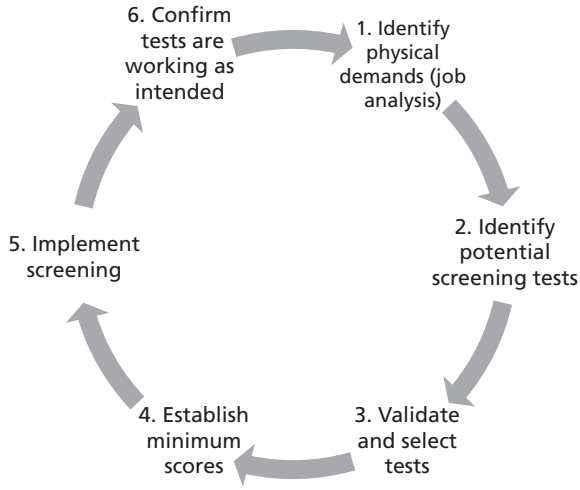
a valid, gender-neutral, and unbiased standard could still result in different proportions of women being selected than men. In other words, finding that a test results in adverse impact against one group does not necessarily mean a test is unfair or biased. Instead, it is considered standard practice to explore more carefully whether score differences are occurring because of differential validity or bias in prediction of later performance or because of real performance differences also observed on the job. Although exploration of differential validity and bias is recommended in all selection contexts, it is especially relevant when selection rates by gender are likely to be unequal, as would be the case with physical standards for ground combat occupations. As a result, to implement the NDAA's mandate to ensure validity and equitable treatment, bias and differential validity by gender also should be explored.

To summarize, the services were directed to establish physical standards that are directly tied to the physical capabilities required to perform the job, are the same for men and women, and do not inequitably screen out members of one gender who are, in fact, able to perform the job. Therefore, the challenge for the military services was to identify standards for each occupation that are valid and equitable in predicting occupation-specific job performance for both sexes and to provide empirical evidence demonstrating that—all by the deadline stated in the NDAA.

Analytic Framework

To assist the military services in developing general and occupation-specific standards relevant to performance, we provide an overview of the process recommended for developing those standards, described in greater detail in Volume 1 of this study. This recommended process is based on the research literature and best practice in other organizations with physically demanding jobs, such as police and firefighters (Hardison, Hosek, and Bird, 2018). The recommended approach involves six stages. As shown in Figure S.1, the six stages are

Figure S.1
Six Stages in Developing Physical Standards



RAND *RR1340/2-S.1*

1. **Identify physical demands (job analysis).** The process for establishing an accurate accounting of a job's tasks or activities that take place in a job is known as job analysis. The job analysis, which is used to design an appropriate selection system, should identify and describe in detail the physically demanding tasks the applicants would need to perform in the job.
2. **Identify potential screening tests.** The second stage is to identify the potential tests that might be used to screen job applicants. Many factors weigh into this decision, but one important consideration is whether research and theoretical support exist for a tool's use in a similar employment context. Other factors include fidelity to the job, cost, and feasibility.
3. **Validate and select tests.** The third stage in developing physical standards is to validate potential tests and identify those with the highest validity and least adverse impact against certain groups. The ultimate goal of validation is to provide evidence that the selection test predicts important outcomes on the job.

4. **Establish minimum scores.** The goal in this stage is to determine the minimum test score(s) that corresponds to acceptable on-the-job performance. Test scores should be anchored to a concrete level of performance.
5. **Implement screening.** When the previous stages have been completed and clear instructions for the proper test administration procedures have been devised, it is appropriate to begin using the screening tool in personnel selection.
6. **Confirm tests are working as intended.** Established standards for entry into physically demanding occupations will need to be the subject of ongoing research to regularly confirm that tests are working as intended and will need to be updated regularly for tests as occupations evolve over time.

We used the stages as a guide for evaluating the methodologies in use by the military services to set gender-neutral standards for most of the ground combat occupations under consideration. With the exception of Air Force battlefield airmen, initial selection of special operations forces personnel is made using a top-down process in which senior personnel rank applicants based on their physical screening test scores and other information. For the top-down selection processes, Stage 4 (establish minimum scores) does not apply. The fourth stage applies only when tests are used to determine whether individuals have met a set of minimum standards (i.e., passing or failing), such as when minimum scores are required for graduation from training or continuation in the occupation.

Because the services were still in the process of developing standards when the collection of information for this research ended in April 2015, our evaluation focused on the first three stages summarized in Figure 3.1 (see Chapter Three) and, to a limited extent, the fourth stage. To understand the activities being undertaken, we met with representatives involved in the research in each of the services, reviewed documentation they provided summarizing the details of the work, and observed some of the data collection efforts.

Evaluating the Service Efforts to Develop Physical Standards

In 2012, 21 percent of the U.S. Department of Defense's (DoD's) active component positions were closed to women—more than 250,000 out of 1.2 million FY 2011 positions (DoD, 2012). From 2012 to January 2016, a number of occupations were opened to women. The positions that remained closed at the end of 2015 were not evenly distributed across the services. As a result, the magnitude of the challenges that the military services faced as they put in place the elements necessary to open remaining closed positions to women differed substantially among the four military services.

The majority of the closed positions were in the Army and Marine Corps—the services with substantial numbers of personnel in the ground combat specialties. In contrast, the Air Force and Navy each had only a handful of positions still closed to women under the ground combat exclusion policy, all of which are among the special operations forces. These special operations occupations (found in all four services) have small numbers of personnel and therefore account for a smaller number of positions relative to the entire force. As with similar positions in the Army and Marine Corps, the special operations positions in the Air Force and Navy were opened to women in 2016.

Just as the numbers and types of closed positions were unique to each service, so too were their efforts to establish standards for those positions. In the following sections and Table S.1, we provide highlights of the findings from our review of those efforts. Technical terms are defined in Appendix A.

Army Combat Arms

Our evaluation of the Army's process for combat arms is based on the combat engineer specialty—one of seven specialties in the combat arms and the only one completed through Stage 3 at the time of our review. The Army's process to validate a set of occupational entry tests included three major data collection efforts. The first effort generally accords with our recommended Stage 1 (conducting a job analysis), whereas the second two efforts most closely accord with Stages 2 and 3

Table S.1
Summary of Key Features of the Service Approaches

Service	Selection Process Being Validated	Stage 1 Identify Physical Demands (Job Analysis)	Stage 2 Identify Potential Screening Tests	Stage 3 Validate and Select Tests
Army Combat Arms	Screening before training	Review of existing job-analysis materials through subject-matter expert (SME) interviews, focus groups, and incumbent survey to rate frequency, importance, time spent	12 candidate predictor tests, chosen to measure types of physical abilities identified by SMEs as needed for physically demanding tasks	Concurrent criterion-related validation to determine how well candidate tests predicted performance on simulated job tasks
Army Special Operations Forces	Training	New in-depth job analysis by the Office of Personnel Management (OPM) using occupational information, site visits, job incumbent survey	Existing training activities	Content validity, details to be determined
Marine Corps Combat Arms (Phase 1 study)	Screening before training	Job tasks identified from existing training and readiness manuals, which rely on occupation-specific task lists regularly updated based on SME review and a job incumbent survey	Elements of existing Physical Fitness Test (PFT) and Combat Fitness Test (CFT)	Concurrent criterion-related validation to determine how well candidate tests predict performance on basic physical tests roughly similar to physically demanding job tasks

Table S.1—Continued

Service	Selection Process Being Validated	Stage 1 Identify Physical Demands (Job Analysis)	Stage 2 Identify Potential Screening Tests	Stage 3 Validate and Select Tests
Marine Corps Combat Arms (Phase 2 study)	Not clear how results will be used to set standards	Unit mission events developed by SMEs representing multiple Marine Corps organizations, including operational combat organizations	Data collected included an unknown number of potential screening tests	Concurrent criterion-related validation to determine how gender mix of a unit and individual physical characteristics affected unit performance and, to a lesser extent, individual performance during unit events
Marine Corps Special Operations Forces	Training	New in-depth job analysis by OPM using occupational information, site visits, job incumbent survey	Existing training activities	To be determined
Navy Special Operations Forces	Training	New job analysis with SME input and job incumbent survey; also developed mission scenarios using focus groups of experienced job incumbents and incumbent survey to determine difficulty, importance, frequency of mission scenarios	Existing training activities (Hell Week in particular)	Content validity through job incumbent judgments of attributes relevant to success in mission scenarios and relevance of Hell Week to actual operations, identified through survey of job incumbents

Table S.1—Continued

Service	Selection Process Being Validated	Stage 1 Identify Physical Demands (Job Analysis)	Stage 2 Identify Potential Screening Tests	Stage 3 Validate and Select Tests
Air Force Battlefield Airmen	Screening before training	Job analysis with review of existing task lists by SME focus groups and survey of job incumbents, and final review by panel of senior and junior incumbents	Identified new tests based on test criteria determined in the research literature, pilot study of 60 candidate tests	Concurrent criterion-related validation to determine how well candidate tests predicted performance on simulated job tasks

(selecting and validating selection tests chosen based on the data collected and the research literature).

The first effort was aimed at defining and evaluating the critical physically demanding tasks in each specialty. This was accomplished by first reviewing existing training activities, field manuals, and task lists to create a preliminary list of physically demanding tasks for each specialty, then revising the list through focus groups with SMEs and a survey of job incumbents.

The second effort involved administering realistic simulations of the most critical physically demanding tasks to help identify candidate selection tests and develop a simplified set of simulations for inclusion in the third effort (a validation study). The realistic simulations were administered to male job incumbents and female volunteers who spent two weeks learning about and practicing the tasks prior to participating in the simulations. The effort resulted in a set of four simplified simulated tasks that were reviewed and approved by a panel of SMEs for use in the combat engineer validation study.

The third effort was a concurrent criterion-related validation study in which data were collected using the candidate selection tests and the simplified simulation activities described above. The study activities were designed, measured, and analyzed with care and attention to

detail. Approximately 150 participants were recruited for the criterion-related validation study; researchers used regression analysis to determine the best predictor tests to include in the selection test battery.

The Army's approach generally aligned with recommended practices and had many strengths. First, the approach included steps to ensure that linkages between key pieces of the work have been demonstrated. Second, the documentation provided sensible and understandable rationales for key study decisions. A third strength was the collection of information from multiple sources throughout the effort. Some gaps remain, however. One is the lack of examination of bias in the testing by gender; whether the tests predicted equally well for men and women is unknown. In addition, the concurrent-validation method used establishes only the relationships between predictor tests and simulated task performance when the data are collected at approximately the same time. It was not clear when the actual selection tests would be administered by the Army during the period of initial entry into military service, and this could have a significant effect on how the minimum standards should be set. In addition, there was no information available, at the time of our review, on the Army's intended protocols for establishing score minimums for entry into the occupations.

Army Special Operations Forces

U.S. Army Special Operations Command (USASOC) regularly reevaluates training standards for special operations forces, but it initiated a new effort to validate these training standards in response to the congressional requirement and enlisted assistance from OPM and the Naval Health Research Center (NHRC). We obtained descriptions of the OPM and NHRC proposed work, but documentation of the actual work was not available in time for our review.

OPM was asked to conduct a new in-depth job analysis for the Special Forces occupation and for service in Ranger units to determine the knowledge, skills, and abilities (KSAs) required in these occupations; the proposed job analysis approach included reviewing background occupational information, as well as conducting site visits and administering a survey. OPM's approach also included exploring whether personnel assessments and standards are based on competen-

cies required for that position. According to USASOC, OPM planned to use statistical techniques to determine the degree to which test activities in the existing training are aligned with tasks identified in the job analysis and are operationally relevant and not unfairly discriminatory; details of these statistical methods were not available during our evaluation.

It appeared that USASOC would be relying heavily on the job analysis work by OPM to provide evidence of the link between the physical training activities and the physical requirements of the special operations jobs. This approach to setting standards reflected the important role of training in screening out candidates for the occupation.

OPM has a long history of conducting job analysis and validation of selection practices using approaches that are generally consistent with recommended practices. That said, some questions remain about this work. More specifically, while there are minimum requirements for applying, there are no formal criteria for selection into the training pipeline. Instead, similar to most of the special operations forces, selection into the training programs is made using applicant rankings by senior USASOC personnel. Because of this top-down selection, the selected applicants typically have standard physical fitness test scores well above the minimums required to submit an application. However, there is no official standardized process for how applicants should be rank ordered. Given that research has shown that unstructured subjective judgments often are not as good as other more-structured and validated approaches to selection, validation evidence supporting the ranking process and its criteria should be collected. However, it is not clear how the information provided by OPM's job analysis could be used to assess the applicant ranking process or determine which training tests are most valid, what the minimum scores on the tests should be, and whether the tests are biased against any relevant groups. Additional data collection and analysis would be required to validate the applicant selection process and training criteria (our recommended Stage 3). Moreover, the OPM job analysis as it was described to us did not explicitly include any plans to consider alternative screening methods beyond those already in place. As a result, we cannot say how well our recommended Stage 2 was being addressed by USASOC's

approach. Lastly, our evaluation of the OPM methodology was based solely on the office's stated plans and past reputation for expertise in this area, not on a summary of the results. Without examination of the actual analyses and findings, our conclusions about the soundness of OPM's methodology are tentative at best.

Marine Corps Combat Arms

The Marine Corps conducted two major studies relevant for developing physical standards for its closed ground combat occupations. The first study explored the correlation between scores on the existing Marine Corps fitness tests and simulated individual physical task performance. To start, the study identified the individual-level physical tasks required of personnel in each occupation and the performance standards for successful completion of these tasks. This information was then used to design a study that correlated the Marine Corps' existing PFT and CFT scores with performance on simulation tasks serving as proxies for the most physically demanding tasks identified in the first study. NHRC then analyzed the data after the Marine Corps collected it. Based on those findings, the Navy researchers recommended a set of screening tests and minimum qualifying scores for selection into these occupations.

Although the process followed in the Marine Corps' first study generally aligned with our recommended stages for developing standards, we identified several limitations in the data and analyses that could affect the soundness of the findings. For example, although the participants in this Marine Corps study included both men and women, the study did not report relationships separately by gender. It is, however, important to note that combining male and female data in the same statistical analyses without controlling for gender can yield misleading results. One possible such result is that a statistical relationship between a predictor test and later performance may disappear when each of the genders is examined separately. Given that nearly all of the men in the Marine Corps study successfully completed all of the proxy tasks, the variance for the female participants and the difference in performance between men and women drove the relationship between test scores and the proxy performance tasks.

Also, the decision to rely on existing fitness tests was made before the physical job tasks were identified, and other potential tests were not explored. In addition, it is unclear whether the simulations developed to measure job task performance could serve as realistic proxies for the actual physical job tasks. Lastly, the findings from the proxy simulations were intended to apply to all closed occupations regardless of whether the occupation required the task it was intended to simulate, and no adjustments for differences in task difficulty across jobs were made.

The second study involved creation of the Ground Combat Element Integrated Task Force (GCEITF) to evaluate the performance of gender-integrated ground combat teams. The task force consisted of 376 Marine volunteers (including 77 women who successfully completed training), who were evaluated as they rotated through a series of simulated mission events. This study was in progress when our evaluation ended.

The GCEITF had the potential to provide additional data and analysis addressing the limitations of the first study. The Marine Corps designed the experiment primarily to determine whether assigning women who successfully complete training to ground combat units affects unit performance, not to develop physical selection standards. Nevertheless, the individual-level test score data and individual-level performance outcomes being collected could support analyses other than those described in the research protocol, including the validation of screening tests and the setting of minimum standards on those tests. Such analyses have the potential to strengthen and supplement the information resulting from the first study. However, we note that our assessment is based only on the design and analytical plans for the experiment. Without seeing the actual data, methods, and results, we could not fully evaluate how useful the experiment was for this purpose.

Marine Corps Special Operations Forces

The Marine Corps Forces Special Operations Command (MARSOC) outlined two principal steps in its approach to establishing standards: (1) conduct a detailed job analysis for the special operations positions of

interest to identify the tasks and abilities required on the job; (2) validate standards in the training courses for these positions. The approach included identifying selection factors and screening tests that determine which trainees are allowed to continue; collecting trainee performance data; and using a hybrid content/criterion-based validation approach to evaluate how well the screening tests predict who can successfully execute the job duties identified in the job analysis (although MARSOC anticipated there might not be time to complete criterion validation work prior to the 2015 deadline). Similar to the USASOC's approach, these efforts were solely directed at validating the selection during the training courses. At the end of our data collection, no plans were in place to validate the processes used for assigning scores to applicants for top-down selection into training, which relies on rankings by senior MARSOC personnel of applicant packages.

MARSOC also contracted with OPM to execute a validation plan. Because of the timing of OPM's contract initiation, our evaluation is limited to the scope of work OPM provided to MARSOC. Similar to the USASOC work plan, OPM's description of the planned job analysis is generally consistent with recommended practice. Other steps beyond the job analysis, however, were laid out in less detail, making it difficult to judge whether the results would provide sufficient support for the physical standards underlying MARSOC's selection process during and prior to entry into training. Details on the tests that would be identified for validation, the type of data that would be collected and analyzed in the validation process, and information about the process that would be used to establish minimum standards were not yet available. Because of the lack of available documentation, there were large gaps in the information we had for assessing the OPM work for MARSOC. Without examination of those actual analyses and findings, our conclusions about the soundness of OPM's methodology are tentative at best.

Navy Special Operations Forces

In the Navy, members of the Sea Air and Land (SEAL) Teams and Special Warfare Combatant-Craft Crewman (SWCC, also known as Special Warfare Boat Operators) were the only closed ground combat

occupations. The Naval Special Warfare Command's (NSWC's) data collection efforts for validating selection standards did not include a reexamination of the physical test requirements for screening candidates into training or for continuation in training; instead, they relied on previous research, some of which was conducted for this purpose. However, there are many gaps in the past research that still need to be filled to provide support for continued use of these testing requirements—including analyzing whether the tests would predict equally well for both male and female applicants, something past studies did not explore.

Instead, similar to the Army and Marine Corps special operations forces, NSWC's effort focused on investigating the extent to which SEAL and SWCC selection requirements that occur *during* training are related to occupational performance. NHRC led that effort. Our evaluation of NHRC's effort was based largely on a draft write-up of its intended methodology. To update an older job analysis, NHRC relied on the input of SMEs and a survey of job incumbents to identify realistic tasks that occur during typical missions—an approach generally consistent with the type of information collected in typical job analysis settings. In addition to identifying tasks, job incumbents were asked to identify physical and personality attributes important to the job—a task that aligns with Stage 2 of the recommended process and an approach that would be strengthened if outside experts provided an independent assessment that agreed with the job incumbent results.

NHRC also included survey questions asking job incumbents to provide judgments about whether physical training activities during the SEAL training Hell Week were important preparation for mission success (an identical approach was used to validate the SWCC equivalent to Hell Week, called The Tour). These key training events screen out a substantial number of trainees. Based on the results of these surveys, the NHRC concluded that Hell Week is valid preparation for serving as a SEAL. However, job incumbent judgment provides a limited basis for such a claim. Collecting other, more objective data to support the link between Hell Week success and actual on-the-job performance would go a long way toward supporting the use of Hell Week as it now stands.

Moreover, none of the data collected by NHRC included women because there were none currently on the job or in training, so it is unclear whether the perceived training-performance relationships would be the same for women and men. As noted earlier, examination of such within-gender relationships is critical to ensuring that a test is valid and unbiased against those subgroups. This is an area that should be explored further in the future.

Air Force Battlefield Airmen

In the Air Force, only seven occupations—as well as the associated units and training courses—were still closed to women because of the ground combat exclusion policy. Personnel in these occupations (both officers and enlisted) are collectively known as *battlefield airmen*. The Air Force effort began with a detailed job analysis to define the critical physically demanding tasks in each job based on information gained from focus groups with SMEs and surveys of airmen in the specialty. In the next step in the process, the Air Force conducted an extensive data collection effort in which a range of physical tests were identified as potential predictors of job performance and then administered to a sample of approximately 200 personnel in various simulations. The results of this validation study were designed to establish training entry requirements and for use in establishing annual testing standards that battlefield airman operators must meet to continue in their occupations. Once the operator tests were selected and minimum test standards set, the researchers planned to complete one more final check of the minimum test scores by having experienced operators complete the tests and then execute full mission profiles as part of the existing operator practice events regularly conducted in the United States. This step would serve to verify that the established standards are working as intended in an operational environment.

In general, the Air Force approach to setting physical standards is consistent with recommended practice—from the job analysis, to identifying screening tests, to many elements of the criterion-related validation effort. The researchers have taken steps to collect solid data on which to base their decisions at important points in the validation process, and their plan included collecting data to confirm many of the

critical links in a well-designed criterion validation study. However, the formal write-up of the methods, analyses, and findings was not available during our study period, so many details of the data analysis could not be considered in our evaluation. In addition, although there are many strengths to the approach that lend credibility and support to any resulting test score minimums, there are some potential gaps in the work. In particular, the plans did not include analysis to test whether the standards would exhibit gender bias.

Overarching Observations, Findings, and Recommendations

Comparing the Service Efforts

The services took different approaches to amassing evidence to develop and support their screening standards. Differences in the approaches do not mean that one effort is better than the others, as there are always multiple sound options for how to approach the work. Nevertheless, those differences will have bearing on what conclusions can be drawn from each of the respective efforts. We highlight several notable differences across the services' efforts in this section and in Table S.1.

Operationalizing "physical screening." Each service conceived of its physical screening in a slightly different way, and, as a result, the work to validate the physical screening processes varied in focus. The Army and Marine Corps work for ground combat occupations and the Air Force's efforts for its special operations occupations were designed specifically to establish gender-neutral standards for selection into these occupations at entry. In contrast, the work by the Army, Navy, and Marine Corps for their special operations occupations focused most heavily on validating the training content and did not validate the use of physical screening test results in the top-down process used for initial applicant selection. Although the information obtained through the research is useful for informing the validity of the other screening elements, the researchers designing the Army's combat arms effort and the Air Force's battlefield airmen effort generally pursued methodologies that were well suited to addressing the first four stages of

our recommended six-stage process. Their investigations relating to each step were carefully designed to provide sound data and solid links between each element of the research and each step. As a result, based on our review of the work they completed by the end of our study, these two services' efforts generally provided greater empirical support for the validity and utility of their screening processes relative to the other efforts described in this report. In sum, while the rest of the services' efforts marshaled some form of support for their screening process, much of that support was—to varying degrees—less definitive than that of the Army's combat arms effort and Air Force's battlefield airmen effort.

We also noted some other differences in the services' approaches:

- **Comparing highly similar jobs across services.** Differences in the services' efforts are likely to receive close scrutiny for jobs that appear to be highly similar across services. Infantry jobs, for example, will be a natural comparison to make between the Army and Marine Corps efforts. The two services have taken very different approaches to collecting and analyzing data for infantry and could well end up with valid but different screening criteria. If these differences affect outcomes for women in a measurable way, they may need to be reconciled with attention to any legitimate reasons for why the differences exist.
- **Establishing occupation-specific versus combat arms-specific standards.** The Marine Corps is the only service that designed a study to establish a single standard for all its ground combat occupations—a reflection of the fact that Marines can be called on to perform duties in any of the combat arms occupations. The other services, however, have not taken such an approach. They have established standards for each occupation that are specific and applicable to that occupation only. If the Marine Corps implements a single screening standard for entry into multiple combat occupations, there inevitably will be differences in the Marine Corps and Army standards that again may require further investigation.

Remaining Stages of the Standard-Setting Process

No single research effort can address all issues, and no research study is without weaknesses and gaps in the analyses and information provided. As a result, Stage 6 (ongoing research to confirm tests are working as intended) will be an important next step after the standards are in place. A number of issues are common to all of the services' work in support of standards for the closed occupations.

- **Our evaluation did not address the service processes for establishing minimum acceptable scores.** At the point of completing our research, none of the services had established minimum selection standards or determined how they would do this. This step is critical to determining whether the standards are set appropriately so that the services are admitting people who are capable of performing on the job, excluding people who are not capable of performing on the job, and not excluding valid candidates unfairly.
- **The physical screening tests used in top-down selection of applicants for Army, Navy, and Marine Corps special operations forces training should be validated.** This effort should also make the rationale for top-down selection explicit and assess the minimum test scores required to apply.
- **The implementation step still needs to be investigated.** Many things could occur during implementation that could invalidate the screening for predicting who will be successful. The services should continue to monitor their implementation procedures to ensure they are being followed and no unanticipated changes (e.g., in test procedures) have occurred that could result in reduced validity.
- **The validity of the standards for female applicants, trainees, and job incumbents should be established in the future.** There was no pool of incumbent women in the closed ground combat occupations with the same experiences or training as their male counterparts for the researchers to draw upon as participants in the research. This was an unavoidable dilemma. To explore whether the standards developed prior to opening the occupations are equally valid for women, the services will need to con-

tinue collecting data on the validity of the screening criteria and alternative measures on samples of both men and women applicants and incumbents in the years following the opening of the positions.

- **Future research may show needed changes.** The services' physical standards will be based on the evidence amassed so far, but more research ultimately will be needed to fully determine how well the tests and test minimums work in practice (Stage 6). Factors discoverable with follow-on research include whether the standard is set too low, whether other tests could be useful as additions or replacements, whether test administration problems arise, retest reliability, and more. It is likely that future research will show that the services should make adjustments and refinements to the selection processes—a normal and necessary part of the process.
- **Formal documentation of all aspects of the work is needed.** All the research we reviewed for this study now has been documented and made public. Similarly, research showing how the final screening tests and minimum scores are chosen and how well the tests perform when they are implemented should also be documented and the documents made public.

Final Thoughts

The call to develop valid standards was taken seriously by the services. All of the services dedicated a large amount of time and resources to their work in response to the lifting of the ground combat exclusion policy. Some services have involved large numbers of voluntary participants (men and women), and some have set aside dedicated testing locations, simulation equipment, and scientific physiological measurement equipment. Although the types of expertise and experience levels of service personnel involved in the efforts differed across the services' efforts, all have sought to involve personnel with a background and expertise in physiological research or personnel selection. Some services

had such experts in house, whereas others sought out the assistance of experts outside of their organizations. The numbers of voluntary participants joining in the work have also been impressive. All told, the work that the services put forth reflects a valiant effort to accomplish exactly what was being requested: the establishment of gender-neutral valid physical standards. That said, some of the services' efforts were more comprehensive, had fewer limitations to the findings, and produced stronger evidence to support the validity of their tests. The fact that the services put forth significant effort and resources should not overshadow the fact that some of the research efforts left more questions unanswered than others.

Acknowledgments

We would like to thank Rennie Vasquez and Lt Col Robert J. Jackson, our project officers during the second phase of this project in the Office of Officer and Enlisted Personnel Management within the Office of the Secretary of Defense. We are also indebted to Juliet Beyler, who directed the office during the period when this phase was conducted, for her guidance and support.

We conducted numerous interviews during this study, and we are indebted to those who shared their time and expertise. Key among these were personnel involved in conducting the work to identify and validate physical standards for ground combat occupations at the U.S. Army Training and Doctrine Command, U.S. Army Research Institute of Environmental Medicine, U.S. Army Special Operations Command, U.S. Marine Corps Training and Education Command, Naval Health Research Center, U.S. Marine Corps Operational Test and Evaluation Activity, U.S. Marine Corps Forces Special Operations Command, and U.S. Air Force Air Education Training Command.

We also wish to thank RAND colleagues who contributed to this effort. Chloe Bird, our coauthor on the Phase 1 report, participated in the early stages of this second phase. Barbara Bicksler drafted the summary of this report. Finally, we received valuable suggestions that helped us improve the report from our technical reviewers Curt Gilroy, Bryan Hallmark, and Neil Carey.

Abbreviations

AAV	assault amphibious vehicle
AETC	Air Education Training Command
AFQT	Armed Forces Qualification Test
AFECD	Air Force Enlisted Classification Directory
AFOCD	Air Force Officer Classification Directory
AFS	Air Force specialties
AIRR	aviation rescue swimmer
AIT	Advanced Individual Training
ANOVA	analysis of variance
APFT	Army Physical Fitness Test
ARSOF	Army Special Operation Forces
ASVAB	Armed Services Vocational Battery
BCT	Basic Combat Training
BFV	Bradley Fighting Vehicle
BUD/S	Basic Underwater Demolition/SEAL
CCT	combat control team
CFT	Combat Fitness Test

CRO	combat rescue officer
CSO	critical skills operators
C-SORT	Computerized Special Operations Resilience Test
DGCAR	Direct Ground Combat and Assignment Rule
EOD	explosive ordinance disposal
DoD	U.S. Department of Defense
GCEITF	Ground Combat Element Integrated Task Force
HPP	Human Performance Program
IRB	Institutional Review Board
IST	Initial Strength Test
ITC	Individual Training Course
KSAs	knowledge, skills, and abilities
MARSOC	Marine Corps Forces Special Operations Command
MCOTEA	Marine Corps Operational Test and Evaluation Activity
MEPS	Military Entrance Processing Station
METL	Mission Essential Task List
MOS	military occupational specialty
ND	Navy diver
NDAA	National Defense Authorization Act
NHRC	Naval Health Research Center

NSW	Naval Special Warfare
NSWC	Naval Special Warfare Command
ODA	Operational Detachment Alpha
O*NET	Occupational Information Network
OPM	Office of Personnel Management
OSD	Office of the Secretary of Defense
PAST	Physical Ability and Stamina Test
PFT	Physical Fitness Test
PJ	pararescue
POI	program of instruction
PST	physical screening test
RASP	Ranger Assessment and Selection Program
SAT	strength aptitude test
SEAL	sea, air, and land (Special Warfare Operator)
SF	Special Forces
SFAS	Special Forces Assessment and Selection
SFG	Special Forces Group
SME	subject-matter expert
SOCOM	Special Operations Command
SOCS	special operations capabilities specialists
SOCS-S	special operations combat service specialists
SOO	special operations officers

SOPC	Special Operations Preparation Course
SOW	statement of work
SOWT	special operations weather team
SQT	SEAL Qualification Training
STO	special tactics officer
SWCC	special warfare combatant-craft crewmen
TACP	tactical air control party
T&R	training and readiness
TECOM	Training and Education Command
TRADOC	Army Training and Doctrine Command
TRMG	Ground Training and Readiness Manual Group
USARIEM	U.S. Army Institute of Environmental Medicine
USASOC	U.S. Army Special Operations Command

CHAPTER ONE

Introduction

The role of women in the U.S. military has been gradually expanding since World War II. Over much of this period, however, women have been precluded from serving in specialties and assignments that involve direct combat on the ground. In the mid-1990s and then more than a decade and a half later, changes in combat-related restrictions on the women in uniform began to take shape. In 1994, then–Secretary of Defense Les Aspin rescinded the "risk rule"—the policy adopted by U.S. Department of Defense (DoD) in 1988 that "excluded women from noncombat units or missions if the risks of exposure to direct combat, hostile fire or capture were equal to or greater than the risk in the units they supported" (Burrelli, 2013).

The change in policy meant that women could be assigned to any position for which they were qualified, with the exception of "those units below the brigade level whose primary mission is to engage in direct combat on the ground" (Aspin, 1994). Although many new positions became open to women when the risk rule was rescinded, the exception, known as the Direct Ground Combat and Assignment Rule (DGCAR), continued to prohibit assignment to occupational specialties or positions in or collocated with direct ground combat units below the brigade level, in long-range reconnaissance and special operations forces, and in positions that include physically demanding tasks the "vast majority" of women cannot do (Aspin, 1994).

Changes in the battlefield environment were one primary motivator in this policy evolution. The battlefield was no longer linear, with a dangerous "front" and comparatively safe "rear." In the 1990s, the

nonlinear battlefield emerged in which military camps and operating bases were surrounded by hostile territory, placing everyone at risk. The wars in Iraq and Afghanistan set the stage for other changes on the battlefield, with women increasingly integrated into military operations. While not assigned to combat units, women participated in combat missions—they flew combat operations, served within range of enemy artillery, interacted frequently with direct ground combat units as part of support units, were exposed to enemy hostilities, and substituted for men in closed positions.

Recognizing this evolution, the fiscal year (FY) 2011 National Defense Authorization Act (NDAA) required review of all laws, policies, and regulations restricting the equitable service of women in the military. This review identified the ground combat rule "as the primary policy restricting the service of female members in the U.S. Armed Forces" (Office of the Under Secretary of Defense for Personnel and Readiness, 2012). In 2012, DoD rescinded the colocation restriction, opening 14,000 combat-support positions to women. Then, on January 24, 2013, almost two decades after the ban was put in place, Secretary of Defense Leon Panetta announced the decision to rescind the 1994 Direct Ground Combat Definition and Assignment Rule (or "ground combat exclusion policy") for the intention to "integrate women into occupational fields to the maximum extent possible" as of January 2016 (Panetta, 2013). In announcing the decision to eliminate the rule, Secretary Panetta stated:

> Our purpose is to ensure that the mission is carried out by the best qualified and the most capable service members, regardless of gender and regardless of creed and beliefs. If members of our military can meet the qualifications for a job—and let me be clear, I'm not talking about reducing the qualifications for the job—if they can meet the qualifications for the job, then they should have the right to serve, regardless of creed or color or gender or sexual orientation (Panetta, 2013).

As the military opened new positions to women, particularly positions with physically demanding tasks, the services needed a more systematic way to determine who would be qualified to fill these posi-

tions. Section 543 of the NDAA for 1994 mandated gender-neutral occupational standards to qualify individuals for any military occupation open to men and women and gender-neutral "specific physical requirements" for open occupations in which performance depends on "muscular strength and endurance and cardiovascular capacity." The FY 2015 NDAA requires that the "gender-neutral occupational standards being developed by the secretaries of the military departments (1) accurately predict performance of actual, regular, and recurring duties of a military occupation; and (2) are applied equitably to measure individual capabilities" (Pub. L. 113-291, 2014). These gender-neutral standards are to be developed, reviewed, and validated no later than September 2015, as specified in Section 524 of the FY 2014 NDAA (Pub. L. 113-66, 2013). And the Secretary of Defense is responsible for ensuring that the standards are developed and implemented according to the statutory requirements.

Mindful of these responsibilities, the Office of the Under Secretary of Defense for Personnel and Readiness asked RAND to help it understand how to evaluate job-specific physical requirements and establish gender-neutral standards for physically demanding jobs. Our study addressed two research objectives. The first objective was to describe best-practice methodologies for establishing gender-neutral standards for physically demanding jobs, tailored to address the needs of the military. Volume 1 of this report documents our work directed toward this objective (Hardison, Hosek, and Bird, 2018). The second objective of the study was to review and evaluate methodologies being used by the military services to set gender-neutral standards.[1] This volume of our report provides the results of this work, using the best-practice methodology established in the first phase of our research as a framework. The review began in 2013, as the services planned their efforts, and ended in April 2015, about six months prior to the 2015 deadline and before the services had completed the final stages of their work. The report focuses on physical standards for military occupa-

[1] See the services implementation plans: Mabus, 2013a, 2013b; Donley, 2013; McHugh, 2013; McRaven, 2013.

tions closed to women. Appendix B addresses physical standards for physically demanding occupations open to women.

Throughout this report, we use the terms *standards* or *physical standards* to refer to occupation-specific criteria that applicants must meet to enter or remain in a particular career field or specialty. We are concerned with standards used to make selection decisions—that is, decisions made that may exclude people from entering or continuing in a job. With respect to gender-neutral standards, we use a simple and straightforward definition. If the minimum passing score is the same for women as it is for men, then it is gender neutral. Nevertheless, according to the NDAA, the intention is not merely that a physical selection standard be the same for both genders, but also that the scores on screening tests be valid (i.e., useful in predicting success in critical physical requirements of the job) and applied equitably.

Exploring impacts on subgroups is one important way to ensure validity and equitable treatment in selection practices. In particular, test validity should not differ among relevant subgroups (such as gender and race), and test scores should be *unbiased* (i.e., two people who receive the same test score should have the same likelihood of success on the job, regardless of subgroup).[2] As a result, it is considered standard practice when establishing physical standards to explore whether differential validity and bias exist. Although this is recommended in all selection contexts, it is especially relevant when selection rates by gender are likely to be unequal, as would be the case with physical standards. As a result, to implement the NDAA's mandate to ensure validity and equitable treatment, bias and differential validity by gender also need to be explored.

[2] A valid, gender-neutral, and unbiased standard could still result in different proportions of women being selected than men. For example, if a job requires lifting 80-pound boxes, asking candidates to pass a test requiring them to lift 80-pound boxes would likely be an unbiased standard for entry into the occupation. That is, it would likely predict who would be successful equally well for men and women. However, it still would likely result in differences in the proportion of women who pass the test and are selected relative to the proportion of men. On the other hand, a test requiring them to lift 150-pound boxes might unfairly exclude many candidates who would be successful in the job, particularly women. This would be an example of a biased gender-neutral standard.

To summarize, the services were directed to establish physical standards that are directly tied to the physical capabilities required to perform the job, the same for men and women, and equitably screen members of both genders. To the maximum extent possible, the standards should let in individuals of both genders with the capability to perform in the job and screen out those who will not perform well. Thus, the challenge for the military services was to identify a single set of standards that is valid and equitable in predicting job performance for both sexes, and to provide empirical evidence demonstrating that, by the deadline stated in the NDAA.

Setting the Stage

In 2012, 21 percent of the department's active component positions were closed to women—more than 250,000 out of 1.2 million FY 2011 positions (DoD, 2012). Since that time, a number of occupations have been opened to women. The remaining closed positions were not evenly distributed across the services. As a result, the magnitude of the challenges that the military services faced as they put in place the elements necessary to open remaining closed positions to women differed substantially among the four DoD military services.

The overwhelming majority of the closed positions can be found in the Army and Marine Corps—the services with substantial numbers of personnel in the ground combat, special operations, and security forces operational specialties. In contrast, the Air Force and Navy each had only a handful of positions still closed to women under the ground combat exclusion policy, all of which are among the elite special operations forces. These special operations occupations have small numbers of personnel. Therefore, they account for a smaller number of positions relative to the entire force. As with similar positions in the other services, the decision to open these special operations positions in the Air Force and Navy to women was made in late 2015.

The Importance of Establishing Valid Physical Standards in Light of the Policy Change

It is worth noting that in some closed occupations, gender alone may have been used as a proxy for whether or not someone would be capable of meeting the physical demands of the job. That is, it was assumed that if you were a man, you were physically capable. However, that assumption would necessarily be wrong for some men. In the past, these men would have been allowed into an occupation even though they would not be likely to meet the physical requirements (i.e, they would have been a false positive). A well-designed physical screener could have done a better job of selection even if only men were allowed into the occupation. Now that gender cannot be used as a proxy to screen people for physically demanding occupations, however, there would likely be even more false positives selected if no valid physical screeners are put in place. Therefore, standards for these jobs need to be established in place of the old gender proxy. Doing so will ensure not only that women and men who are incapable of performing the job are screened out but also that women who are capable of doing the job well are not overlooked. The purpose of finding a more valid screening tool is always to find a more-capable force, which necessarily will result in a greater likelihood of mission success. By opening up jobs to women, DoD acknowledged that using gender as a proxy for physical capability is doing a disservice to the military by screening in male false positives (men who would not be successful on the job), while screening out female false negatives (women who would be successful on the job).

In some of the closed occupations, physical screening standards have been in place for years; however, the validity of the screening criteria may not have been established using accepted best practices, or the process used to validate them may not have been well documented. In those cases, the validity of the screening criteria could be questioned. The NDAA language applies to these cases as well, by explicitly directing the services to demonstrate that the existing standards are valid for predicting who will be successful.

Even if studies have demonstrated validity in the existing job incumbent population, it is not clear whether that validity will hold when the jobs are open to a wider range of applicants. Validity on the

standards would likely have been explored only for men, as applicants and job incumbents could only be male. Now, with the applicant pool changing, it is possible that screening criteria that were ideal for a male-only population will not be suitable when the population is broadened. More and even entirely different criteria may be needed.[3] Thus the NDAA also recognizes that any screening criteria previously validated on a male-only sample necessarily need to be reexamined to ensure that they are still valid when used on the broader population.

Lastly, although demonstrating that the services have established valid standards is especially important in light of the policy change opening up combat jobs to women, doing so on a regular basis would still have value absent any change to DGCAR. That is, even in a male-only population, a failure to validate existing standards could lead to a less-effective force. Therefore, it is recommended practice for all selection criteria to be validated periodically to reduce the number of false positives and false negatives in a selection process and to ensure that selection practices truly result in improved success of those on the job. In the private sector, doing so is recommended to improve an organization's bottom line; in the military, it is recommended to improve mission success. In that sense, the NDAA's mandate to establish valid standards could have occurred at any time, but the issue of validating standards is now in the spotlight because of the change to the DGCAR policy.

Organization of This Report

Our report begins in Chapter Two with a description of the best-practice methodology for establishing gender-neutral standards for physically demanding jobs—a construct used in the remainder of the report

[3] First, testing both genders is the only way to find out whether the test is valid for both groups. By exploring validity on a wider sample, tests with the greatest validity for both groups and the least adverse impact can be identified. Second, if testing only one gender, the wrong conclusion could be drawn about the importance of including the test as a predictor. Suppose all men could lift an 80-pound box, but only 50 percent of women could do so. In testing only men, someone might decide the test is unnecessary, when in fact it might prove to be an important predictor of performance in the broader applicant pool that includes women.

to evaluate the work by the military services to set gender-neutral standards. Chapter Three describes our analytic approach. Our evaluations of the services' efforts begin in Chapters Four and Five, which describe our assessment of the Army's activities to examine physically demanding positions in general combat arms and special operations forces, respectively. Chapters Six and Seven examine physically demanding positions in Marine Corps combat arms and special operations forces. Chapter Eight reports on the Navy's special operations forces. Finally, Chapter Nine discusses battlefield airmen—the Air Force's special operations positions that will be newly open to women. Chapter Ten offers a discussion of the similarities and differences in the services' approaches and key aspects of the work still to be completed that will be important for implementing, monitoring, and adjusting the new policy down the road. As noted earlier, we also include two appendixes as supplemental information. Appendix A provides an overview of many of the technical terms used throughout the report. We encourage readers to consult that appendix as needed. Appendix B provides an overview of some of the services' existing screening processes for physically demanding jobs already open to women.

CHAPTER TWO

Recommended Process for Establishing Physical Standards

Civilian employers whose jobs are physically demanding have long faced scrutiny regarding the appropriateness and equity of their standards. DoD can expect similar scrutiny as it embarks on the process of developing gender-neutral physical standards—and, for this reason, wishes to employ appropriate methods in this endeavor. To assist the military services in developing occupation-specific standards relevant to performance, we provided an overview of the process for developing those standards described in Volume 1 of this work based on the research literature and best practice (Hardison, Hosek, and Bird, 2018). We grouped the process into six stages to reflect that it necessarily involves attending carefully to each stage in the process. For each stage, important features are associated with good practice in carrying out the work. We based these features on best practices recommended in the personnel research literature and used by other organizations with physically demanding jobs (such as police and firefighters) that must screen applicants for suitability before they enter these careers.

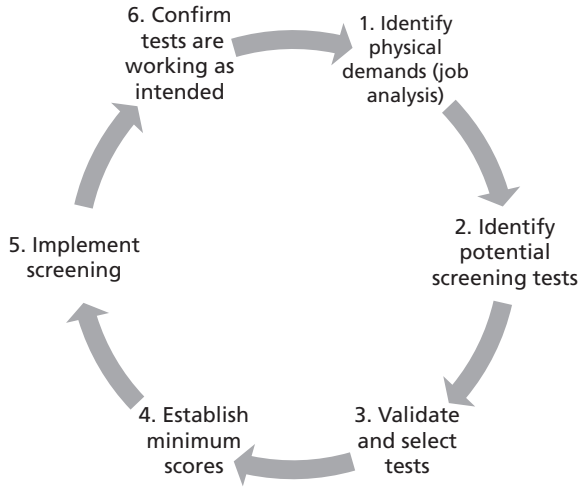
The six general stages for establishing physical job requirements are shown in Figure 2.1. Each stage provides critical support for determining an appropriate set of selection procedures. Carrying out the entire process requires the involvement of researchers with expertise in a variety of domains, including industrial and organizational psychology, exercise physiology or a related field, psychometrics, and statistics. These technical experts also rely on the expertise of subject-matter experts (SMEs) from the occupation, who must be carefully selected

to cover all types of work and work environments, and on appropriate test subjects drawn from the population of applicants, trainees, and job incumbents. The deliberate implementation of each stage and careful documentation of the actions taken are central to developing defensible physical standards.

This six-step process applies to occupations where the goal is to establish minimum standards for initial selection and continuation in training and on the job. Some military occupations, including most of the special operations forces, take the best candidate among those who apply until the available positions have been filled. That process is known as top-down selection. In top-down selection, evidence supporting a robust linear relationship showing that higher scores on the screening tests used are associated with higher levels of performance on the job is necessary to defend the use of test results in top-down selection. This evidence can be developed following Stages 1 through 3 of the process shown in Figure 2.1. As discussed in Hardison, Hosek, and Bird (2018), Stage 4 of the process is not necessary to support top-down selection, but it might be desirable to establish minimum scores for these jobs.

In many cases, top-down selection may not be preferred; in some cases, it may not be defensible. In these cases, establishing minimum scores needed for selection into a job could be a better alternative. A variety of circumstances should lead an organization to consider establishing minimum score standards instead of or in addition to top-down selection. Jobs in which performance below a certain level could pose a danger to others is one example of a circumstance where drawing a minimum performance line would be critical. Moreover, in some jobs, there might be no real benefit to gains in job performance, beyond a certain threshold or performance level, especially at the very high end of the score distribution. Jobs fitting these circumstances are especially good candidates for establishment of minimum physical ability scores, given the potential adverse impact of the screening against women. Any time top-down selection is chosen over establishing minimum standards, a clear rationale for why it is appropriate should be provided, along with evidence that supports that rationale. We further

Figure 2.1
Six Stages in Developing Physical Standards



RAND *RR1340/2-2.1*

discuss top-down selection in the chapters describing setting physical standards for the services' special operations forces.

An overview of the six-stage process, which applies to the other occupations addressed in this report, is provided in this chapter. In addition, for further reference, Appendix A explains many of the key terms used in this and later chapters in the report.

Stage 1. Identify Physical Demands (Job Analysis)

The process for establishing an accurate accounting of the tasks or activities that take place in a job is known as job analysis. The results of a job analysis serve as the foundation for nearly all types of human resource management activities, to include an organization's selection system. Job analyses can be conducted in several different ways. Some are worker-oriented approaches that focus on what workers do in performing their jobs; others are job-oriented approaches that focus on what workers accomplish in their jobs. Both approaches are valid and

result in the collection of distinctly different types of information. Choosing among these alternatives, as well as determining how data are collected and what experts are called on to assist in the process, should be driven by the goals for the job analysis.

In establishing gender-neutral requirements for entry into physically demanding jobs, the focus is on applicant selection. Therefore, the goal of the job analysis should be to design an appropriate selection system. It should identify and describe in detail the physically demanding tasks the applicants would need to be able to perform in the job. In this context, task-level detail that is specific to the particular occupation under study is ideal for a sound defense of a selection system. It is also important to ensure that SMEs and others involved in the job analysis have adequate experience and sufficiently represent the overall worker population—to include relevant representation among employment locations and varying seniority of personnel who undertake the work. If performed correctly, the results of the job analysis should set the groundwork for other stages in the process of establishing requirements. If a job analysis is to be used for a different purpose (e.g., continuation in a job), it needs to be designed with that purpose in mind. Similar issues arise in setting standards for continuing in a job, but the focus would be on testing job incumbents instead of applicants.

If a job analysis has recently been done for an occupation for which standards are being established and/or validated, it should be carefully reviewed to ensure that its description of the physical demands is complete, accurate, and sufficiently detailed to support the remaining steps in the standard setting process.

Stage 2. Identify Potential Screening Tests

The next step in developing physical standards is identifying potential tests that might be used to screen job applicants (or job incumbents). In this context, we use *screening* to refer to evaluation of individuals' physical skills relevant for performing the job tasks described in Stage 1. Many factors weigh into this decision, but one important consideration is whether research and theoretical support exist for a tool's

use in a similar employment context. Test developers and employers should be aware of relevant research results—whether new tests are being explored or well-established tests are being considered.

Selecting the right tests in an employment context requires careful attention to which physical abilities are and are not required by the job. Once these are determined, a variety of factors come into play when selecting a test: fidelity to the job, cost, and feasibility are three of the most important. Fidelity to the job refers to the similarity between the test and job tasks. High-fidelity tests have obvious overlap with the job and are often viewed as more fair by test-takers. Low-fidelity tests have little observable similarity to job tasks but instead measure general physical abilities that could be relied on to perform job tasks. There can be some overlap in the two types of tests, and either type or a combination of both can be used effectively to screen job applicants.

Cost and feasibility are closely aligned and often relevant in choosing between high- and low-fidelity tests. All relevant costs must be considered, to include equipment costs, manpower costs, and validation costs. Feasibility relates to how difficult it is to implement a test in multiple locations. Cost and feasibility are of particular concern to the military services in, for example, considering whether to scale up an occupation-specific test for use by recruiters. Furthermore, because the military has many different physically demanding jobs, it faces unique challenges in selecting a set of tests for initial job classification. Using high-fidelity tests, in this context, may well be cost-prohibitive. Instead, administering a series of simple tests that can generalize across multiple jobs might be a more feasible approach.

Where physical standards already exist for the occupation, the test(s) already used should be included in the list of tests to be considered. To guard against the possibility that standards based on these tests prove not to be valid, other potential tests should also be considered.

Stage 3. Validate and Select Tests

The third stage in developing physical standards is to validate potential tests and identify those with the highest validity and least adverse

impact. In the personnel selection context, the term validate has a precise meaning. It refers to the act of accumulating multiple sources of research-based evidence to support a test's use for a particular purpose. The ultimate goal of validation in a personnel selection context is to provide evidence that the selection test predicts important outcomes on the job.

Best practice requires evidence be accumulated to support claims that a test measures what it is intended to measure and that its scores can be used for selection. There are various types of validation evidence that an organization can collect, and each piece of evidence lends additional support to that claim. Validation evidence helps to answer several questions: Does the test fully capture the relevant characteristics of the physical requirements? Is there a clear relationship between test scores and outcome measures? Do the outcome measures capture important job outcomes? If tests are deficient, then candidates might be selected who are not capable of performing on the job or candidates might be screened out who would be capable.

Collecting validation evidence is a complex process. When undertaking validation studies, an organization must document all aspects of the research study design and its results. These studies typically require considerable statistical and methodological expertise and a careful design before data collection begins to ensure that results are statistically sound (enough statistical power, representative of the population, etc.). In addition, bias against key subgroups (e.g., women or minorities) should be explored, which requires statistical oversampling of those subgroups. Finally, organizations should seek multiple sources of validation evidence whenever possible.

Stage 4. Establish Minimum Scores

The next stage in the process is to establish the minimum scores that reflect acceptable performance on the job. The goal in this stage is to determine the minimum test scores that correspond to acceptable on-the-job performance. Test scores should be anchored to a concrete level of performance, such as lifting a certain number of pounds or running

a specific distance within a certain amount of time. Minimum scores should be set consistent with Secretary Panetta's commitment to not "reducing the qualifications for the job" in announcing the elimination of the ground combat exclusion rule:

> Our purpose is to ensure that the mission is carried out by the best qualified and the most capable servicemembers, regardless of gender and regardless of creed and beliefs. If members of our military can meet the qualifications for a job—and let me be clear, I'm not talking about reducing the qualifications for the job—if they can meet the qualifications for the job, then they should have the right to serve, regardless of creed or color or gender or sexual orientation (Panetta, 2013).

The process of establishing minimum cutoff scores, referred to as standard-setting, is distinct from validation. When used in an employment context, it typically involves convening panels of experts to identify the test score that distinguishes a competent performer from one who is not competent. (In some cases, it might be possible to rely on job analysis data to justify a minimum score.) But because all experts may not agree, best practice requires a systematic approach that solicits the perspectives of a variety of people. The ultimate goal of standard-setting is to make the resulting minimum cutoff score as objective and reliable as possible. Thus, documenting the process by which the score is established is also critical.

Stage 5. Implement Screening

Once the previous stages have been completed and clear instructions for the proper test administration procedures devised, it is appropriate to use the screening tool in personnel selection. But a number of key issues should be addressed during the implementation stage to ensure that the test is implemented in a manner consistent with the results of the validation and standard-setting efforts.

The timing of test administration can influence results. Tests administered far in advance of the predicted work should have evi-

dence to show that the time gap does not change the validity of the test or the interpretation of the test scores. For example, basic training is an event that would be expected to improve all applicants' physical abilities. Tests administered in advance of basic training could underpredict performance for everyone unless training effects are accurately taken into account—something that should be included in the validation process. It is also important to standardize test administration procedures so that each person has an equal opportunity to demonstrate his or her capability on the test regardless of where it is being administered. Key to standardization is creating clear documentation of the proper administration procedures and ensuring the equipment and testing environment are the same at all test locations.

Other important factors during implementation include informing applicants about the test so they have an equal opportunity to prepare. In addition, when new tests are instituted, an organization may want to phase in the test so that applicants have enough time to become familiar with the test and prepare for it. Phasing in tests also allows an organization to collect additional data to further validate the test in an operational setting.

Stage 6. Confirm Tests Are Working as Intended

Once initial standards for entry into physically demanding occupations are established, they will need ongoing research to regularly confirm that tests are working as intended. Even the best research designs leave some questions unanswered. New, unanticipated questions may arise after implementation. Some studies are feasible only after a test has been implemented. Changing technology and missions can significantly alter the requirements of the job. And new research findings may emerge that suggest changes in testing policies. For all these reasons, the research effort should be treated as an ongoing process—one that continues long after a test has been implemented. Ideally, research efforts examining all stages of the standard-setting or validation process would be institutionalized as part of a regular operational data-collection activity for each occupation—a process not new to the military services.

Summary

The methods for establishing physical standards for specific occupations involve this six-stage process. The first four stages contribute to the initial development of the standards—the tests and minimum test scores that will be employed in selecting among applicants for entry into an occupation or among job incumbents for continuation in the job. The tasks conducted in each stage are essential for ensuring that the standards accurately reflect the physically demanding work in an occupation, measure physical capabilities needed to carry out that work, and are set at the right level for successful performance on the job.

Gender-neutral physical standards are set without regard to gender and reflect only the physical capabilities needed to perform tasks associated with the occupation. However, to ensure that standards are not biased against either gender, the process of validating tests and setting minimum test scores must be based on data collected from women and men. When an occupation has been closed to women, the developers of standards must find a pool of women with related training and experience to represent women who might enter the occupation in the future.

Once the standards have been developed, the last two stages of the six-stage process focus on implementation and sustainment. Without careful implementation and ongoing monitoring and updating, even well-designed standards will fail to screen individuals appropriately when the testing is done improperly or as occupational tasks and equipment change over time.

CHAPTER THREE

The Analytic Approach for Evaluating the Services' Efforts

We used the stages and the best-practice methods described in Chapter Two (and more fully in Hardison, Hosek, and Bird, 2018) as a guide for evaluating the methodologies being used by the military services to set gender-neutral standards—which are described in the chapters to follow. Since the services were still in the process of developing standards when our research ended, this evaluation focuses on the first three stages summarized in Figure 3.1 and, to a limited extent, the fourth stage. The last two stages are elements that should be examined (implement screening and confirm tests are working) but that could not be completed prior to the September 2015, the deadline for establishing gender-neutral and valid standards set in Section 524 of the FY 2014 NDAA.[1] Instead, those elements will be next steps for the services' implementation and evaluation of the standards.

Therefore, we structure the discussion of each service's effort around the four stages shown in Figure 3.1. While it is ideal to work through these steps in a deliberate, sequential manner with each step informing the next, each of the services approached the process of setting gender-neutral standards from a different starting point—some having amassed data relevant to elements of the process before the Sec-

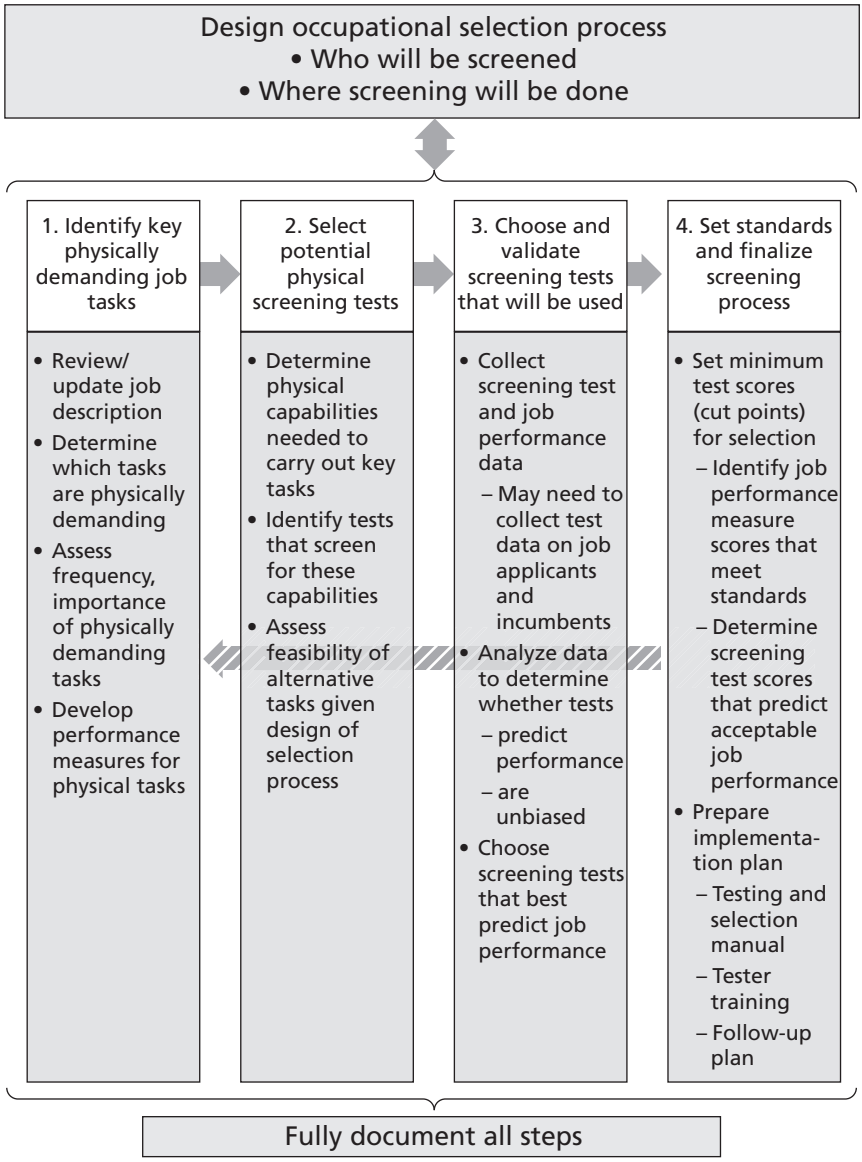
[1] This is a potential catch-22, in which the standards need to be approved before they can be implemented and further tested to ensure they are working as intended (Stages 5 and 6), but approval to implement the standards cannot be issued until valid standards have first been established. The services would legitimately have to stop short of the last two stages because approvals to proceed would be required before they could move on to Stages 5 and 6.

retary of Defense lifted the DGCAR. But by using the four-stage process as an organizational framework for our research, we are better able to determine whether and how well the services' new data collection activities and previously amassed data address the important elements that should be covered at each stage—placing less importance on whether they did them in precise order. The reverse arrow from right to left in Figure 3.1 indicates that the process may not be strictly linear; results of later stages can make it necessary to return to earlier stages. For example, if none of the tests selected in Stage 2 achieves an acceptable level of validity in Stage 3, it may be necessary to redo Stage 2 with other tests.

To understand the activities being undertaken, we met with representatives involved in the research in each of the services, reviewed documentation they provided summarizing the details of the work, and observed some of the data collection efforts. The representatives we sought out were those most knowledgeable about the details of the methodology. This typically included some discussion with organizational representatives assigned with the responsibility of overseeing the work, plus extended discussion with the researchers who were actually conducting the study and collecting and analyzing the data. Some of the research documentation they provided to us was unpublished—such as human subjects research protocols and study materials submitted to their Institutional Review Boards (IRBs); draft technical reports summarizing methods, data analyses, and findings; internal briefing slides; and memos. Final technical reports on the work conducted were subsequently published on a DoD website (in December 2015, after our work was completed). We also relied on published and unpublished work conducted in prior years to support occupational standards established before DGCAR was rescinded.

Discussions took place over several years, starting in late 2012 when the services were just beginning their research efforts. In those early meetings, the services' work plans were in their infancy, so we met with the services periodically to learn more about their details as the research unfolded. The summaries provided in the following chapters present the culmination of those discussions.

Figure 3.1
Physical Standards Development Process



RAND *RR1340/2-3.1*

As the work progressed over the multiyear period, the services made adjustments and improvements to their plans. Such changes were expected for two reasons. First, some of the services had teams of in-house personnel who were experienced at conducting this type of research effort, whereas others needed to establish teams and seek out the assistance of external organizations. Given this, some of the initial descriptions of service plans were highly detailed at the very beginning of the effort, whereas others were only conceptual, with few details provided to us on how exactly the studies would be implemented at that point. Second, even highly detailed research plans can change as the work progresses—analyses should be driven by the data that is collected, and methodologies should be adjusted depending on pilot data findings, for example. Making such changes is always necessary in a research study, as many relevant details and issues in the design cannot be adequately addressed until after the research has been initiated.[2] Over the course of the project, we observed changes in all of the service efforts as they progressed. As a result, only the final details about the work completed or planned as of April 2015 are documented in this report.

The various discussions with each service took place by phone, video teleconference and/or in-person. The questions were unstructured, but we provided a list of topics ahead of time and started by asking the services to walk us through each step in their study design and probed for additional details about important features of the design in each part of the research. The following are examples of questions asked about each stage of the research, as well as about how the services were documenting their efforts. Exact questions, however, depended on the specifics of the research in question.

[2] Information obtained after beginning a study (such as pilot study data or sample constraints) can sometimes lead to significant changes in the research approach.

- Stage 1. Identify physical demands (job analysis)
  - How have you defined the physical demands of the job? Was there a job analysis? What was the process?
  - Who participated? Did you use SMEs? How were they selected?
  - How many individuals? How many groups? How were the results analyzed?
- Stage 2. Identify potential screening tests
  - What tests have you considered for possible use? On what basis?
- Stage 3. Validate and select tests
  - What process was used for validation (e.g., predictive validity, content validity, convergent and discriminant validity of the test) and why?
  - For predictive validation, what outcomes are you measuring in the study? Who was used for the sample? How were they selected and why? What statistical analyses are you using to evaluate the results?
  - Did the sample include women?
  - Have you examined whether there are differences in the predictive validity of the tests by gender or any other groups or whether the test underpredicts performance of any group?
- Stage 4. Establish minimum scores
  - How are you establishing test score minimums? Describe the process.
  - When will the test be administered?
  - Would people be expected to improve on the test between the time in which the test will be administered and the time at which they will be expected to be proficient?[3] Have you conducted a study estimating the amount of improvement expected (e.g., as a result of basic training or technical training)?

[3] Declines in performance could also occur, for example, because of time in the delayed entry program. If declines in performance are expected among individuals waiting to enlist, research could also be conducted to estimate the magnitude. However, it is unclear whether raising standards in anticipation of expected declines in performance would be fair, particularly when greater adverse impact could occur, and the declines could be preventable. Instead, that information would ideally be used to establish new training and testing interventions during the intervening periods to prevent the declines from occurring.

- Documentation
  - What supporting documentation do you have or plan to have summarizing the work in each stage?

At the conclusion of our data collection period in April 2015, none of the services had completed its work. None had arrived at the end of Stage 4, establishing minimum scores for selection, or provided a description of how these would be established. In addition, the Marine Corps, Army, and Navy had not yet established clear policy regarding the point in someone's tenure (prior to joining, upon joining, after boot camp, etc.) when the screening tests will be administered.[4] The timing of the administration of the testing will matter for setting minimum test standards in Stage 4. Given that the Stage 4 work was not complete, we cannot provide a detailed overview of the services' processes for establishing the final minimum test score cut points that will determine whether men and women can enter or continue in a specific occupation.

Only in the case of the Marine Corps has any analysis been completed and documented that explicitly addresses minimum scores; however, that work will likely be combined with the results of another major ongoing study to finalize the minimum scores. Therefore, with the exception of the Marine Corps, the following chapters focus primarily on Stages 1 through 3; we discuss the implications for the Stage 4 work more generally in Chapter Ten.

---

[4] The Air Force has specified that force-wide screening tests will be administered at the Military Entrance Processing Station (MEPS), and battlefield airman tests will be administered at several times starting as early as recruiting stations (for more on this, see Chapter Nine).

CHAPTER FOUR

Army Combat Arms

In addition to the Special Forces (SF) occupation and Ranger assignments discussed in Chapter Five, four Army combat arms branches were closed to women at the time of this study. In total, this accounted for more than 110,000 officer and enlisted positions in the Army, as of May 2013. About 10,000 are enlisted positions in the Engineer branch, and nearly 15,000 are enlisted positions in the Field Artillery branch. The overwhelming majority, however, are found in the Armor (around 27,000 enlisted and more than 2,000 officer positions) and Infantry branches (about 57,000 enlisted and more than 3,000 officer positions).

The Army's Training and Doctrine Command (TRADOC) has had primary responsibility for the work to establish gender-neutral standards for these closed positions. To accomplish this, TRADOC tasked the U.S. Army Institute of Environmental Medicine (USARIEM) with developing and implementing the methodology for establishing physical performance screening requirements for entry into the following seven closed military occupational specialties (MOSs):

- 11B Infantryman
- 11C Infantryman-Indirect Fire
- 12B Combat Engineer
- 13B Cannon Crewmember
- 13F Fire Support
- 19D Cavalry Scout
- 19K Armor Crewman.

USARIEM staggered the work such that data collection for each MOS had a different start date. As a result, the work was still underway for the majority of the MOSs when we completed data collection. However, the work on the combat engineers, the first MOS to be undertaken, was complete by this time. Thus, our description of the Army's process later in this chapter will focus primarily on combat engineers, with the understanding that this is the process the Army planned to use for each MOS in turn.

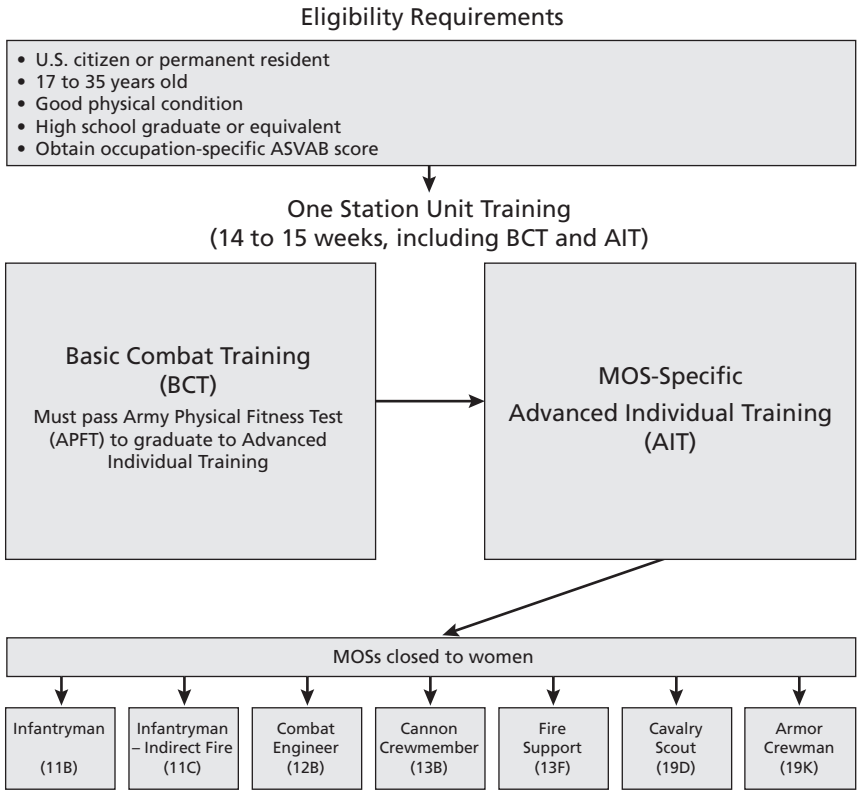
Occupational Assignment and Screening in the Army

Army recruits sign an enlistment contract specifying the occupation they will enter at their local MEPS. The eligibility requirements for closed occupations are the standard requirements all Army enlistees must satisfy—e.g., holding U.S. citizenship or permanent residency, being ages 17 to 35, exhibiting good physical condition and moral standing, graduating from high school or holding an equivalent certification, and scoring above a stated Armed Services Vocational Battery (ASVAB) threshold for that occupation (Figure 4.1). Career counselors at the MEPS offer recruits a choice of MOS assignments based on their ASVAB scores and the Army's current personnel needs.

Candidates who fulfill the eligibility requirements enter One Station Unit Training, which includes seven-week Basic Combat Training (BCT), followed by a four- to five-week occupation-specific Advanced Individual Training (AIT). To graduate from BCT to AIT, recruits must pass the Army Physical Fitness Test (APFT).

The APFT is administered to all Army personnel, regardless of whether the occupation is closed to women. It includes three scored events: push-ups, sit-ups, and a 2-mile run. Scoring standards specify optimum (100 percent) and minimum (60 percent) thresholds that vary by gender and age group for the push-up and run events and vary only by age for the sit-up event. For example, the optimum push-up score for 17-to-21-year-old men is 71, and the optimum for 17-to-21-year-old women is 42. If candidates do not pass the APFT toward the end of BCT, they will continue in a longer BCT cycle, with more opportuni-

Figure 4.1
Eligibility and Training Path for Enlisted Personnel in Army Closed Occupations



SOURCE: Headquarters, Department of the Army, undated.
RAND *RR1340/2-4.1*

ties to take and pass the APFT. Candidates who do not pass after an extended time period (i.e., six months) are typically discharged.

Army Pamphlet 611-21 provides a detailed description of every Army MOS (Headquarters, Department of the Army, undated). Each description includes a list of all physical tasks in the job and the frequency of each. Table 4.1 shows the physical tasks for Infantry, which includes those in the 11B MOS.

Table 4.1
Physical Tasks for Infantry (11B)

Physical Tasks
1. Frequently visually identifies vehicles and equipment at 1,000 meters and individuals at 300 meters.
2. Occasionally drags 271-pound person 15 meters.
3. Constantly performs all other tasks while carrying a minimum of 80 pounds, evenly distributed over entire body.
4. Frequently digs, lifts, and shovels 11 pounds [sic] scoops of dirt in bent, stooped or kneeling position.
5. Frequently hears, gives, or echoes oral commands in outside area up to 50 meters.
6. Frequently walks, runs, crawls, and climbs over varying terrain and altitude changes for a distance of up to 15 miles, during a 24-hour period, while carrying 103 pounds evenly distributed over entire body, after which soldier must retain the ability to perform all other physical requirements.
7. Frequently rise from a prone, kneeling, or crouched position, sprint for 3 to 5 seconds while carrying a minimum of 80 pounds, evenly distributed over entire body, then returning to a prone, kneeling, or crouched position. Repeating for a distance of no less than 100 meters.
8. Occasionally lifts 107 pounds 5 feet high as part of a two-soldier team.
9. Occasionally lifts, lowers, and moves laterally 59 pounds 3 feet while seated.
10. Frequently lifts and lowers 40-pound bags shoulder high.
11. Frequently throws 1-pound object 35 meters.
12. Frequently lifts 45 pounds waist high and carries it up to 15 meters.
13. Occasionally lifts 65 pounds vertically 5 feet to 6 feet in the air.
14. Frequently lifts 65 pounds 3 feet high, moves laterally 5 feet and places object in tube.
15. Occasionally carries 153 pounds 10 meters as part of a two-soldier team.
16. Frequently scales and climbs over a 2-meter vertical obstacle, with assistance.
17. Occasionally raises a 207-pound person 3 1/2 feet as a member of a two-soldier team.

SOURCE: Table 10-11B-1 in Headquarters, Department of the Army, undated.

Army's Process for Establishing Standards for Combat Arms

The Army's process to validate a set of occupational entry tests included three major data collection efforts. The first was aimed at defining and evaluating the critical physically demanding tasks in each MOS. The second involved administering simulations of the critical physically demanding tasks to help identify potential screening tests and develop a simplified set of simulations for inclusion in a criterion-related validation study. The third was a concurrent criterion-related validation data collection effort using the candidate tests and simplified set of simula-

tion activities. The first step is roughly aligned with our Stage 1 (conducting a job analysis), whereas the second two steps most closely align with Stages 2 and 3 (identifying and validating the selection tests).

USARIEM conducted each effort separately for each MOS beginning with combat engineers, the MOS completed and documented during our study period. Therefore, the sections that follow describe the process and preliminary findings for combat engineers. It was the Army's plan to follow the same steps for the other six MOSs; however, we note that the analytic details for each occupation likely differed to some degree.

The following sections describe the steps in the Army approach. The description is pulled from a variety of sources of information, including interviews with the researchers, observations of elements of their data collection efforts, review of unpublished IRB protocols for various elements of the research design, and a preliminary draft of a technical report summarizing the results for the combat engineers.

Identify Physical Demands

The process to identify physically demanding tasks began with a review of existing training activities, field manuals, and task lists for each MOS, conducted by TRADOC with assistance from USARIEM. From this material, TRADOC identified an initial list of physically demanding activities typical for each MOS and created a description containing the details needed to simulate the activity. Many of the identified tasks applied to more than one MOS. A combined list of the tasks identified across all MOSs is presented in Table 4.2. This initial task list served as the starting point for the simulations administered to participants in the second data collection effort, described in the next section.

The simulation descriptions provided in the Army's IRB protocols are listed in Table 4.2. The following are paraphrased descriptions for a few of the tasks (Sharp, 2014a):

- Conduct a tactical movement. Soldiers complete a 12-mile movement while wearing approximately 103 pounds of equipment

Table 4.2
Initial Task List Used in the Army Simulation Observation Study

Occupational Related Task	MOS
Conduct a tactical movement	All
Prepare a fighting position (fill and emplace sandbags)	All
Drag a casualty to immediate safety (dismounted)	All
Remove a casualty from a wheeled vehicle (mounted)	11B, 19D, 13F, 12B
Lift, carry, and install the barrel of a 25-mm gun on Bradley Fighting Vehicle (BFV)	11B, 19D, 13F, 12B
Remove the feeder assembly of a 25-mm gun on BFV	11B, 19D, 13F, 12B
Load 25-mm H-EIT Tracer ammunition cans onto BFV	11B, 19D, 13F, 12B
Load TOW missile launcher on BFV	11B, 19D
Move over, through, or around obstacles	11B, 11C
Move under direct fire (3- to 5-second rushes)	11B, 11C
Prepare dismounted TOW firing position	11B
Lift and carry M2 .50-caliber machine gun	11B
Lift and emplace base plate for 120-mm mortar	11C
Lift and emplace cannon for 120-mm mortar	11C
Fire a mortar (lift and hold round, place in tube)	11C
Mount M2 .50-caliber machine gun on Abrams tank	19K
Stow ammunition on Abrams tank	19K
Load the 120-mm main gun on Abrams tank	19K
Remove a casualty from Abrams tank	19K
Transfer ammunition with an M992 Carrier (M795 ME rounds)	13B
Emplace 155-mm Howitzer (lift wheel assembly)	13B
Displace 155-mm Howitzer (lift spade trail arm and blade)	13B
Establish an observation point (carry AN/PED-1[LLDR])	13F

Table 4.2—Continued

Occupational Related Task	MOS
Install and remove Fire Support Sensor System (F3S) on M1200	13F
Carry and emplace the Antipersonnel Obstacle Breaching System	12B
Carry and emplace the H6 40-pound cratering charge	12B
Carry and emplace the modular-pack mine system	12B
Lift and carry rocking roller during construction of Bailey Bridge	12B
Load and install a Volcano mine delivery system	12B

SOURCE: Sharp, 2014a.

(basic uniform, personal protective equipment, and 24-hour sustainment load).

- Move under direct fire (three-to-five-second combat rushes). While wearing an 83-pound fighting load and carrying a weapon, soldiers start in prone position. On command, they rise, sprint to the first marker 20 meters away and assume a kneeling position. After five-second pauses between each activity, they execute the remainder of the activities: rise and sprint to a second marker 20 meters away and assume a crouched position; rise and sprint to third marker 15 meters away and assume prone position; rise, and sprint to fourth marker 15 meters away and assume kneeling position; rise and sprint to fifth marker 15 meters away and assume crouched position; rise and sprint to sixth marker 15 meters away and run across the finish line.
- Prepare a fighting position (fill and emplace sandbags). While wearing an 83-pound fighting load, soldiers shovel sand into buckets to simulate filling a sandbag. They complete 26 repetitions. Each repetition equals about 30 to 40 pounds of sand. Soldiers then move 26 sandbags (approximately 40 pounds each) 10 meters, where they build a fighting position that is three sandbags in length and three sandbags in height.

- Drag a casualty to immediate safety (dismounted). Soldiers drag a 270-pound casualty a distance of 15 meters as quickly as possible while wearing an 83-pound fighting load.
- Remove a casualty from a wheeled vehicle (mounted). While wearing the fighting load minus the weapon (approximately 75 pounds), soldiers pull a simulated 207-pound casualty from the commander's seat of a BFV or Striker as quickly as possible. This task is performed individually and by two-person teams.
- Lift, carry, and install the barrel of a 25-mm gun on the BFV. As part of a two-man team and wearing an 83-pound fighting load, soldiers lift and carry the 107-pound barrel of the M242 25-mm gun for the BFV 25 meters and emplace it on the vehicle.
- Move over, through, or around obstacles. While wearing or carrying a fighting load, soldiers scale a 2-meter wall with assistance. Equipment may be removed, but it must still clear the wall.

To further confirm the initial list for each of the MOSs, researchers held focus groups with SMEs at two base locations. One focus group consisted of experienced lower-level job incumbents (those most likely to have experience performing physically demanding tasks on the job). The second involved higher-ranking job incumbents (those most likely to have experience supervising people performing these tasks in real-world environments). Focus group participants were asked to review the initial task list to confirm which tasks were performed in their MOS, to verify the accuracy of the task descriptions, and to determine whether any tasks were missing from the list. If changes appeared necessary from the focus groups, the task list was revised. The revised task list was then sent to TRADOC for final review.

This final MOS-specific task list was then sent as an online survey to all job incumbents. In the survey, respondents were asked to rate the frequency, importance, and time spent on each task. The three items on the survey were combined to create a total score for each task. Results were then used to identify the most important tasks to include in the criterion-related validation data collection effort (described later in this chapter).

Winnow the Job Task Simulation Activities

The next step in the Army's process was to winnow down the simulation activities into a manageable set for inclusion as outcomes in the criterion-related validation study. To do this, USARIEM sought to better understand the physical demands in each of the simulation activities. The simulations for the nine tasks relevant to combat engineers, as shown in Table 4.2, were administered to 23 male job incumbents and 11 female volunteers (women were recruited from across the Army). The women were similar to the men in average age (24 and 22, respectively) and military tenure (two to three years on average).

The simulations were designed to have high fidelity to the tasks. For example, they included the actual equipment described in the activity (e.g., BFVs were used in the *Remove a Casualty from a Wheeled Vehicle* activity). However, the simulations were also administered in a controlled setting to ensure that potential sources of error were kept to a minimum. For example, soldiers shoveled sand into buckets instead of sandbags to prevent the possibility that sandbag openings would flop over in the midst of filling, which could confound the results.

The male and female participants spent two weeks together learning about the tasks and practicing them as a group prior to participating in the simulations. This training time allowed the women with no prior knowledge of the tasks to learn how to perform the tasks to standard and served as a refresher for the male job incumbents. It also allowed participants to practice as a team for those tasks that required two or more people. During the simulations, USARIEM measured participants' perceptions of exertion (using Borg CR1-10 and 6-20 scales), VO_2 max[1] and self-reported run times[2] for use in calculating VO_2 max, heart rate, completion times, and distances obtained, as appropriate to the task. Participants were also given a questionnaire in which they

---

[1] This is the maximum amount of oxygen the body can use during a specified period of intense exercise and depends on body weight and the strength of the lungs. It is commonly measured by increasing the intensity of exercise on a treadmill or cycle ergometer while measuring oxygen consumption. The name is derived from *V* for volume, O_2 for oxygen, *max* for maximum.

[2] We understood that soldiers were asked to report their run time from their most recent regular PFT.

indicated how often they completed the tasks in the field and in training. Data for combat engineers showed that the job incumbents who had deployed had engaged in some, but not all, of the MOS-specific tasks in the field.

Using data from the simulations, tasks with similar physical demands were grouped into one of four groups according to the results of the physiological measurements obtained during the simulation study described earlier. A set of four simulated tasks was then designed or chosen to be representative of the types of physical demands and activities found in the tasks within that group.

These four simulated tasks were then used as the outcomes to be predicted in the criterion-related validation study. USARIEM selected the four simulated tasks using the following criteria: safest to administer; requires little to no learning or experience to perform; could be performed by an individual rather than a team; represents the activities most frequently performed by personnel in the field; and represents the most physically demanding task required. A complete rationale for choosing each of the four new simulations is provided in the Army's write-up of the study findings.

For example, one new task, *casualty evacuation from a vehicle turret*, was designed to represent the physical activity of heavy lifting found in the *remove a casualty from a wheeled vehicle, carry and emplace the modular-pack mine system, lift and carry rocking roller during construction of bailey bridge,* and *load and install a volcano* tasks. Because the *remove a casualty from a wheeled vehicle* task was determined in the more extensive simulation observation study to be the most physically demanding of the four, the simulation was designed to emulate that task most closely. In that task, soldiers reached down into a BFV and pulled a heavy bag out of the vehicle. The task was modified for use in the criterion validation study to instead involve pulling a heavy bag onto a raised platform from below. It was also modified to start with a 50-pound bag for familiarization and warm-up. During the actual testing, the bag was increased by 10 pounds until it reached 210 pounds or the soldier being tested could not perform the task.

A description of all four abstracted simulation tasks is as follows (Sharp, 2014b):

- *Casualty evacuation from a vehicle turret.* Heavy lifting (physical ability: muscular strength [remove a casualty from vehicle; carry and emplace the modular-pack mine system; lift and carry rocking roller during construction of Bailey Bridge, load and install Volcano mine delivery system]). While wearing a 71-pound fighting load (full fighting load minus a weapon), soldiers squat, grasp the handles of the heavy bag level through a hole in a platform. They then stand and pull the bag through the hole and onto the platform. They first lift 50 pounds. If they are successful, the weight is increased in 10-pound increments up to 210 pounds. Final lift weight is recorded.
- *Prepare a fighting position.* Repetitive lifting and carrying (physical abilities: muscular endurance and aerobic capacity [prepare a fighting position, load 25-mm H-EIT Tracer ammunition cans on the BFV, carry and emplace the H6 Cratering Charge]). While wearing 71 pounds (full fighting load minus a weapon), soldiers carry 16 40-pound sandbags 10 meters, and place them on the floor as quickly as possible. Soldiers are timed and heart rate is recorded.
- *Casualty drag.* Quickly dragging a heavy object (physical ability: power [drag a casualty to immediate safety]). Soldiers drag a simulated 270-pound casualty 15 meters as fast as they can in 30 seconds, while wearing an 83-pound fighting load. If they fail to pull the casualty the appropriate distance within the time allotted, the distance dragged is measured.
- *Tactical foot march.* Load carriage (physical abilities: aerobic capacity, muscular endurance, and muscular strength [conduct a tactical movement, carry and emplace the Antipersonnel Obstacle Breaching System]). Soldiers complete a movement of 4 miles while wearing the basic soldier uniform, personal protective equipment (including weapon), and 24-hour sustainment load (103 pounds). Soldiers complete this task as quickly as possible while walking on a supervised course with breaks as needed. Time to completion, split-times, and heart rate are recorded.

These four tasks were presented to a new SME panel of nine 12B Sergeants First Class and explained in detail. The SME panel was asked to evaluate whether these newly designed activities (including such relevant details as times, weights, distances, standards for performance) were still relevant and appropriate for the job. All agreed that they were.

After the four simulations were confirmed by the SMEs, they were administered to 25 male and 25 female volunteers from a variety of MOSs to measure test-retest reliability of the tasks. The load carriage simulation was administered twice, and the other three simulations were administered four times. Such as factors as heart rate, time to completion, and perceived exertion also were measured. Results suggest that additional instructions might be needed for two of the tasks to prevent the possibility of score increases due to learning effects, but all tasks exceeded their threshold of acceptable test-retest reliability.

Identify Potential Predictor Tests

The method for establishing the link between physical abilities and the tasks listed in Table 4.2 was outlined in the early research protocols. These protocols stated that 25 SMEs selected by TRADOC from each MOS would be asked to rate how many of the various types of physical abilities (e.g., muscular strength, muscular endurance, anaerobic power, trunk strength) are needed to accomplish each task. The questionnaire items were to be pulled from the Occupational Information Network (O*NET), a well-known source of job analysis information sponsored by the U.S. Department of Labor/Employment and Training Administration. The information from the SMEs was then used—along with the simulation observation study information and a review of the research literature—to inform the selection of a set of predictor tests for inclusion in the criterion-related validation study. A total of 12 tests were selected for inclusion. Some would be familiar to most people (e.g., a minute of push-ups, a minute of sit-ups, and a timed 300-meter run), whereas others would be recognized only by those versed in the physical testing research. Examples of the predictor tests include (Sharp, 2014a)

- *Illinois Agility.* The length of the course is 10 meters, and the width (distance between the start and finish points) is 5 meters. Four cones are used to mark the start, finish and the two turning points. Another four cones are placed down the center an equal distance apart. Each cone in the center is spaced 3.3 meters apart. Soldiers will lie prone (head to the start line) and hands by their shoulders. On the "Go" command the stopwatch is started, and the soldier gets up as quickly as possible and runs around the course in the direction indicated, without knocking over the cones, to the finish line, where the timer is stopped.
- *Standing Broad Jump.* Soldiers stand behind a line marked on the ground with feet slightly apart. A two-foot takeoff and landing is used, with swinging of the arms and bending of the knees to provide forward drive. Soldiers attempt to jump as far as possible, landing on both feet without falling backward. Three attempts are allowed.
- *Handgrip.* Soldiers hold the dynamometer in their hand, with the elbow at a right angle and at the side of the body. The handle of the dynamometer is adjusted such that the base rests on first metacarpal (heel of palm), while the handle rests on middle of four fingers. When ready, soldiers will squeeze the dynamometer with maximum isometric effort, which is maintained for about three to five seconds. No other body movement is allowed. Three trials are given for each hand. The highest two trials on each side are averaged.
- *Arm Endurance Test.* The test involves cranking an arm ergometer for two minutes. The workload (i.e., measure of resistance) is fixed at 50 watts. The test is performed with the candidate in a kneeling position facing the arm ergometer with the center crank adjusted to shoulder height. Following a brief warm-up, the soldier rotates the crank arm as rapidly as possible for two minutes. The total number of revolutions and final heart rate are recorded.

Validate the Screening Tests

Approximately 150 participants (male volunteers from the MOS and female volunteers from other MOSs) were recruited for the criterion-related validation—103 were men, and 43 were women. Both groups were on average of similar ages (about 24 years old) and tenures (about three years).

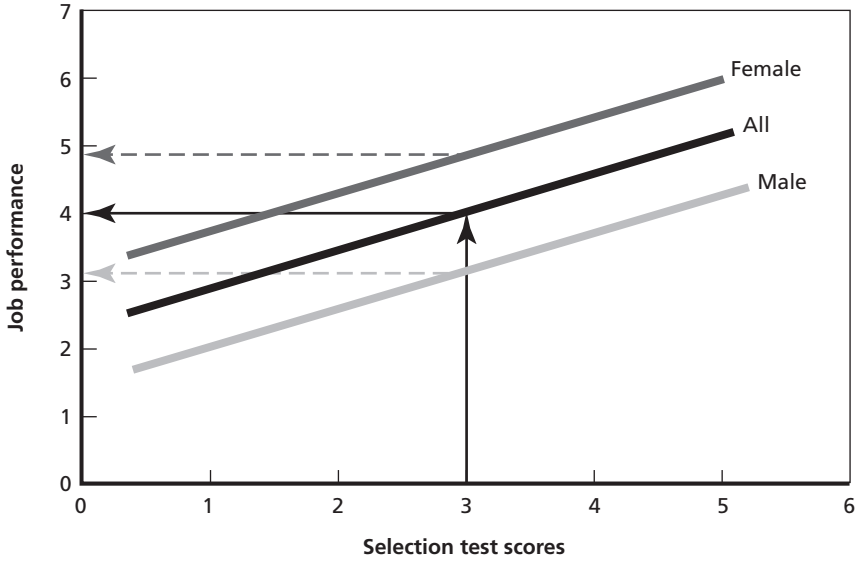
During the study, participants completed the four simulations and all 12 predictor tests. Testing was completed in three different sessions with 24 hours or more between testing sessions. Researchers in the study collected a wide range of data, such as heart rate, perceived exertion, number of repetitions, testing times, and other relevant test scores.

To determine the best tests to include in the selection test battery, the researchers created regression models using the test scores to predict a composite score created from performance on the four simulations. The composite score being predicted was a simple sum of the scores converted to z-score units on each of the four tasks (i.e., the testing time or highest weight achieved, depending on the task). The researchers identified four viable regression equations. The first included only the best predictors. The remaining models had only the best predictors among those that also met other practical criteria, including costliness, ease of administration, and not requiring any specialized equipment. The four equations were then used to predict each of the individual simulation activities, and the resulting correlations were reported. Those correlations ranged from the low 0.60s to the high 0.80s.

USARIEM recommended that the Army select one of the three regression equations for use as the formula for combining information from all the various tests into a single overall selection score. It also recommended conducting additional follow-on criterion-related research on actual selectees to ensure the equations are working as intended in the recruit population and to replicate the results of this work on a new group of participants.

The researchers did not explore whether the regression equations show differential validity, including over- or underprediction by gender

Figure 4.2
Illustration of a Hypothetical Selection Test Biased in Favor of Male Applicants



RAND *RR1340/2-4.2*

(for an illustration of over- and underprediction, see Figure 4.2).[3] Instead, the regression equations were estimated on the pooled results for both male and female participants. In addition, the work to date has not addressed the minimum scores on the screening tests for selection into the occupation. Instead, the researchers acknowledge that no minimums have been established on the simulation activities and that those will need to be established. Establishing such minimums does not appear to be within the USARIEM scope of work.

[3] For more discussion of differential validity and over- and underprediction, see Volume 1 of this study (Hardison, Hosek, and Bird, 2018).

Our Evaluation

Overall, the Army's approach included many elements pertaining to our recommended Stages 1 through 3. The job analysis (Stage 1) work relied on existing documentation of the job requirements and expert judgment provided by TRADOC, but these were reviewed and updated through focus groups with occupation SMEs and information from a survey of job incumbents. USARIEM's study protocols do describe that a survey of all job incumbents was conducted to determine importance and frequency of the tasks identified by TRADOC. However, we did not receive documentation of the job analysis work; therefore, we cannot judge whether the survey findings were consistent with the information provided by TRADOC or how TRADOC arrived at the final task list as provided to USARIEM. The results of the survey and TRADOC's approach are not included in the draft technical report that USARIEM shared with us.

The Army's effort did address our Stage 2, in that it chose a wide variety of screening tools to include in the validation effort. Some screening tools are already in use in other countries (as described in their preliminary technical report). Some were based on the results of SME panels that were asked to judge which physical aptitudes were required to perform the 12 required tasks (as described in their IRB protocols). However, very little rationale for the selection of the final set of tests was offered in the documentation provided to us.

The Army has also completed work in line with recommended practices for our Stage 3, validating the screening tests. Its approach used a simulation-based criterion-related validation study in which the simulations were designed, measured, and analyzed with care and attention to detail.

However, USARIEM's work stopped at the completion of our Stage 3. Information about how the Army would establish minimum scores for entry into the combat engineer occupation (our Stage 4) was not available for our evaluation.

Among the strengths of the Army's work is that the approach takes steps to ensure that linkages between key pieces of the work have been demonstrated. For example, USARIEM-collected data informed

which of the original 12 task simulations could be combined to create four representative tasks for use in the criterion-related validation study. That links the predicted outcomes in the criterion validation study to the original tasks identified by TRADOC. Thoughtful attention to these types of linkages helps lend strength to the end results of the work.

A second strength was the amount of documentation available and in progress. In that documentation, many rationales for key study decisions are provided. For example, the study staff eliminated one task from the list provided by TRADOC because it was more related to practicing the activity than to someone's underlying physical aptitude; in addition, the required physical aptitude to implement the task was low relative to the remaining 12 tasks. The rationale provided for this change is sensible and understandable, lending additional credibility and support to the work by leaving few questions unanswered.

A third strength is the collection of information from multiple sources throughout the effort. For example, after developing the four tasks designed to represent the 12 tasks for combat engineers, USARIEM conducted an SME panel to determine whether the tasks still represented key tasks and capabilities required in the MOS. When multiple sources of evidence provide similar conclusions, the results have stronger support overall.

Although there were strengths to the approach that will lend credibility and support to any resulting test minimums, there are still some areas with gaps. For example, the lack of examination of bias of the testing by gender is one area that was not well addressed in the work. Whether the tests predicted equally well for men and women is unknown. Given the difference in MOS experience by gender among test subjects, it is possible that experience, which was confounded with gender, could account for some of the predictive power of the tests. The limited time provided to practice the tests may not have been sufficient to eliminate the effect of experience in every instance. In addition, the testing established only the relationships at one point in the career (at three years of service). It is not clear when the tests will be administered in the Army (the policy had not yet been decided when USARIEM conducted its research); however, if the tests are administered as early

as enlistment, the minimums established would need to be lowered to account for improvement from training (such as BCT) between the screening point and when soldiers are placed on the job. Finally, the documentation we received did not cover the job analysis, so we were not able to review the methods and data used for this critical first step in the process.

CHAPTER FIVE

Army Special Operations Forces

Army Special Operations Forces (ARSOF) consists of SF, Ranger, Special Operations Aviation, Psychological Operations, Civil Affairs, and Signal and Combat Service Support units. Many of the personnel assigned to these units are in occupations not specific to the ARSOF, and some of these occupations were already open to women. However, assignment to any of the ARSOF units was closed to women at the time of this study. The largest units are the SF and Ranger units, so we discuss entry into these units in more detail in this chapter.

There is a dedicated SF MOS (18X), also known as the Green Berets. Soldiers entering this occupation first complete the training to be an infantryman and then training specific to the SF occupation. According to the U.S. Army Special Operations Command (USASOC) website,

> Special Forces Green Berets deploy and execute nine doctrinal missions: unconventional warfare, foreign internal defense, direct action, counter-insurgency, special reconnaissance, counter terrorism, information operations, counter proliferation of [weapons of mass destruction] WMD, and security force assistance. There are five active component Special Forces Groups [SFGs] and two U.S. Army National Guard Groups. Each SFG is regionally oriented to support one of the war-fighting geographic combatant commanders. The cornerstone of the SFG's capability is the Operational Detachment-Alpha [ODA], a highly trained team of 12 Special Forces Green Berets. Cross-trained in weapons, communications, intelligence, medicine, and engineering, the ODA member also

> possesses specialized language and cultural training. The ODA is capable of conducting the full spectrum of special operations, from building indigenous security forces to identifying and targeting threats to U.S. national interests....The Special Forces Green Berets provide a viable military option for operational requirements that may be inappropriate or infeasible for large conventional forces. (USASOC, undated)

Unlike SF units, the Rangers have no designated MOS. Instead, the 75th Ranger Regiment consists of personnel trained in a number of occupations and who meet the special requirements to qualify for added training required to become a Ranger. Some, but not all, of these occupations (11B Infantryman) are closed to women; others are open to women, but Ranger assignment is closed. In this chapter, we describe selection and training to become a Ranger. Chapter Four focused on selection and training for closed occupations not assigned to Ranger or SF units.

Although there is no Ranger occupation, in general the process of becoming a Ranger resembles the method of entering the SF occupation. Rangers first complete training in a range of occupations instead of all completing infantryman training. USASOC describes the 75th Ranger Regiment on its website as follows:

> The 75th Ranger Regiment is a lethal, agile and flexible force, capable of executing a myriad of complex, joint special operations missions in support of U.S. policy and objectives. Today's Ranger Regiment is the Army's premier raid force. Each of the four geographically dispersed Ranger battalions are always combat ready, mentally and physically tough and prepared to fight the War on Terrorism. Their capabilities include air assault and direct action raids seizing key terrain such as airfields, destroying strategic facilities, and capturing or killing enemies of the Nation. Rangers are capable of conducting squad through regimental size operations using a variety of infiltration techniques including airborne, air assault and ground platforms. (USASOC, undated)

Like the SF, Rangers also work in 12-person ODA teams, with each member of the team contributing a different occupational area

of expertise. SF units account for about 7,000 closed positions and the Rangers for more than 2,000.

USASOC has accepted primary responsibility for establishing gender-neutral standards for the SF MOS and the Ranger Regiment in response to the lifting of DGCAR, and those efforts are described later in this chapter.

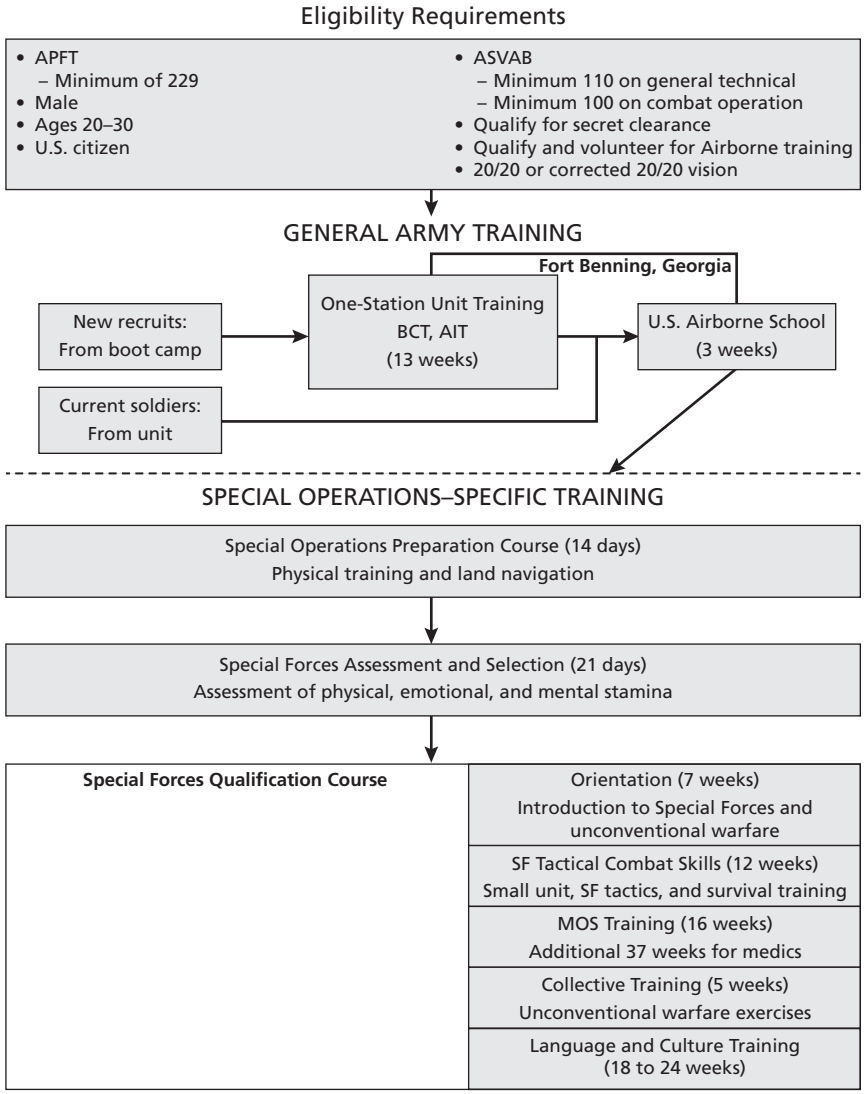
Occupational Assignment and Screening in USASOC

Selection for Army SF training is made using a top-down selection procedure. The process for entering the SF MOS is illustrated in Figure 5.1. The process for Rangers is similar. New recruits, as well as those already in the Army, can apply for entry into the training for these occupations. To be eligible to enter the SF occupation and join the Rangers, applicants must

- achieve passing scores on the APFT, which includes a 2-mile run, push-ups, and sit-ups. Although passing is a requirement, recommended goals for applicants to be competitive to be chosen for training include completing the 2-mile run in 12 to 14 minutes and 80 to 100 sit-ups and push-ups
- have no physical limitations
- be a male, ages 20 to 30 (for SF)
- be a U.S. citizen with a high school diploma
- obtain a general technical score of 110 or higher on the ASVAB for the SF MOS and 105 or higher to join the Rangers; for SF, also have a combat operation score of 100
- qualify for secret clearance
- qualify and volunteer for Airborne training
- for SF, have 20/20 or corrected 20/20 vision.

These are the minimum requirements to submit an application, not to qualify to enter SF training. For most other Army occupations, soldiers are selected for an occupation if they meet the stated requirements and there is a training slot available in their time frame. In con-

Figure 5.1
Eligibility and Training Path for Army Special Forces



SOURCE: Headquarters, Department of the Army, undated.

RAND *RR1340/2-5.1*

trast, selection into the ARSOF is done by senior special operations forces personnel who rank-order applicants based on their holistic assessment of all of the selection criteria (to include physical fitness test scores). Applicants are accepted into the assessment and screening process and subsequently the Special Forces Qualification course based on that ranking and number of seats available. The selection criteria are not explicit; they are implicit in the judgments of the senior personnel rating applicants. There are more applicants than training seats and, not surprisingly, the scores on the APFT for selectees are well above the minimum scores to apply.

Although personnel are prescreened on the physical fitness criteria before they even enter the SF training pipeline (as described earlier), that prescreening is only a small part of the selection process for these occupations. Instead, much of the selection occurs during the training pipeline itself. Each MOS has a training block specifically dedicated to making selection decisions regarding which trainees can continue.

For SF, the first stage of training is 16 weeks of infantry One Station Unit Training, which includes both the BCT course and Advanced Individualized Training course in a specific MOS (Figure 5.1). Upon completion, SF trainees enter U.S. Airborne School/Basic Airborne Training (three weeks, focused on parachuting and landing safely) at Fort Benning, Georgia. The third stage is the first SF-specific training course, the Special Operations Preparation Course (SOPC), which is a 19-day course at Fort Bragg that focuses on developing physical fitness and land navigation skills. After SOPC, recruits begin Special Forces Assessment and Selection (SFAS), the 21-day "survival-training" program, also at Fort Bragg, which is designed as a major selection point where trainees' physical and mental strength is challenged. Activities during SFAS include running, swimming, sit-ups, pull-ups, push-ups, obstacle course, marches, land navigation/orienteering and leadership and teamwork.

For those SF trainees who make it through SFAS, the remainder of training is called the Special Forces Qualification Course. The course is divided into six phases over 62 weeks (with an additional 37 weeks for medics), including

1. Orientation (six weeks): This phase features a full course plus training in unconventional warfare, SF history, organization, task/core activities, capabilities, methods of instruction, patrol orders and troop-leading procedures.
2. Small Unit Tactics (13 weeks): This phase teaches advanced marksmanship, counterinsurgency, urban operations, live fire maneuvers, and sensitive site exploration, as well as survival, evasion, resistance, and escape exercises.
3. MOS Training (15 weeks): This phase provides SF-specific training in the MOS, including language training, SF tasks, advanced special operations techniques, and intra-agency operation.
4. Collective Training/Robin Sage (two weeks): During this simulation, soldiers work on squads on a mission to counter political turmoil in a fictional country that covers vast areas of North Carolina.
5. Language and Culture (24 weeks): This phase offers training in one of the following—French, Indonesian-Bahasa, Spanish, Arabic, Chinese-Mandarin, Czech, Dari, Hungarian, Korean, Pashto, Persian-Farsi, Polish, Russian, Tagalog, Thai, Turkish, or Urdu.
6. Graduation (one week): outprocessing, soldiers now wear green berets as SF soldiers.

The initial training pipeline for Rangers is shorter than for SF (shown in Figure 5.1). Like SF, they first complete BCT and AIT and then Airborne School. After that, they enter the last course required prior to becoming a member of the 75th Ranger Regiment: the Ranger Assessment and Selection Program (RASP). The course is eight weeks, consists of two phases, and is used both to select and screen the trainees and to train them in the fundamentals of the occupation (e.g., marksmanship, mobility, and physical fitness). RASP (like SFAS) is highly physically challenging and accounts for a large part of the screening during Ranger training. Those who pass RASP go on to serve in the Ranger Regiment.

After having served in the Ranger Regiment, typically for a few years, rangers attend 62 days of Ranger School, which is required for personnel assuming higher levels of responsibility in the Ranger Regiment. The Airborne and Ranger Training Brigade website describes Ranger School training as follows:

> Ranger students train to exhaustion, pushing the limits of their minds and bodies. The course incorporates three phases (Benning, Mountain, and Swamp) which follow the crawl, walk, run, and training methodology. In Benning phase, the students become trained on squad operations and focus on ambush and recon missions, patrol base operations, and planning before moving on to platoon operations. In Mountain phase, students develop their skills at the platoon level in order to refine and complete their training in Swamp phase. After these three phases, Ranger Students are proficient in leading squad and platoon dismounted operations around the clock in all climates and terrain. Rangers are better trained, more capable, more resilient, and better prepared to serve and lead Soldiers in their next duty position. (Airborne and Ranger Training Brigade, undated)

Typically, only about 45 percent of recruits complete the entire SF or Ranger training pipeline, with a large portion of the trainee losses happening during SFAS and RASP.[1] The stated purpose of SFAS and RASP is to collect the assessment data that will be used for final selection decisions about who to send to the remainder of training.

Although historically no women have attended SF or Ranger training, USASOC in 2015 opened a number of seats across the January through April Ranger training cohorts to female volunteers. Two women successfully completed RASP in August 2015, but they could not be assigned to the Ranger Regiment pending the decision to open it to women.

---

[1] Only 42 percent of those who attempted Ranger School from 2010 to 2014 completed the course, with the majority of the failures (62 percent) from the Ranger Physical Assessment (Airborne and Ranger Training Brigade, undated). Thirty-six percent of students fail in the first four days.

Army's Process for Establishing Standards for Special Operations Forces

In Chapter Two, we indicated that any screening tests used in top-down selection should be validated by establishing an empirical relationship between higher scores on the tests and performance in training and ultimately on the job. To our knowledge, this has not been done for APFT scores used in selection for Army SF or Ranger training.

Standards for training in USASOC are regularly reevaluated using individual and task-level performance data collected routinely by the Army. As part of the Army's process, course performance metrics are provided to a Critical Mission Task Review Board. The review board conducts a Critical Mission Task Analysis to determine the critical training tasks (individual and unit level) and associated performance metrics. One example of a critical task is a 12-mile march with specified gear and time. No additional information was provided to us regarding the process the Army uses to conduct the Critical Task Analysis.

Consistent with the important role the training programs have for selection of SF and Ranger personnel, USASOC initiated a new effort to validate SFAS and RASP in direct response to the lifting of DGCAR. To accomplish this review, USASOC turned to the Office of Personnel Management (OPM) and the Naval Health Research Center (NHRC) for assistance. The overarching goal for the work (as stated by USASOC) was to establish the relationship between the training tasks required in the courses and those documented in Army Pamphlet 611-21 (Headquarters, Department of the Army, 2007).

Identify Physical Demands

The last job analysis to be completed for the special operations forces was completed by the Army Research Institute in 1998. USASOC asked OPM to complete a new in-depth job analysis for the SF MOS and Ranger Regiment positions. The goal of the research effort was to ultimately determine the knowledge, skills, and abilities (KSAs) required of SF and Ranger soldiers. OPM's job analysis approach

includes reviewing background occupational or positional information, as well as conducting site visits and administering a survey.

The OPM effort was just beginning as we completed our research, so we were not able to acquire details on the methodologies used for its analysis or documentation of the results. USASOC described the anticipated results as including a list of the tasks, competencies, and physical abilities required on the job and anticipated that OPM would provide detailed documentation on the method and results once the work was completed.

Validate the Selection Criteria

According to OPM's scope of work, the final deliverable, due by the end of the third quarter of FY 2015, would be "a comprehensive, documented job analysis that addresses (1) selection and competitive promotions, (2) job-related requirements, and (3) that personnel assessments and standards are based on competencies required for that position." According to USASOC, OPM planned to use statistical techniques to determine the degree to which SFAS and RASP activities measure the KSAs identified in the job analysis, are operationally relevant, and are not unfairly discriminatory. Although USASOC noted that OPM would use statistical techniques to accomplish this, details on those techniques and the data underlying them were not available. We inferred from the information we received that the OPM job analysis work was designed to provide content validity evidence to support the relevance of the training activities and the minimum standards for performance expected in those activities.

USASOC also indicated that it would be providing data to the NHRC exercise physiologists, who would assist in a criterion-related validation and standards validation process (time and data permitting). We interviewed the NHRC researchers at the end of our study period, but they were not able to provide further information on their effort at that time.

Lastly, as noted earlier, selection into the Army special operations forces is done by senior special operations forces personnel who rank-order applicants based on their holistic assessment of all of the selection criteria (including physical fitness test scores). So, unlike most

other occupations, the prescreening process for selection into training involves a great deal of subjective judgment. Given that research has shown that unstructured subjective judgments are often not as good as other more-structured and validated approaches to selection (see, for examples, Dawes and Corrigan, 1974; and Tversky and Kahneman, 1974), research validating these subjective selection decisions should be explored. Although physical fitness scores (along with other information) are considered in the holistic assessment prior to permission to enter training for these occupations, USASOC did not include those assessments in its work in response to the lifting of DGCAR. As a result, at the time in which we completed our interviews, USASOC's work did not include any efforts to validate the process for screening people prior to training.

Our Evaluation

The Special Operations Command (SOCOM) effort to set standards prior to October 2015 was begun relatively late. When we visited SOCOM headquarters in January 2015, work was just being initiated and would not be completed until well after our project work ended. Therefore, we lacked the details necessary to evaluate the reasoning, logic, and methodological soundness of the approach for the validation effort described previously.

From the information provided, it appeared that USASOC planned to rely on the job analysis work by OPM to provide evidence of the link between the physical training activities (particularly those in RASP and SFAS) and the physical requirements of the SF jobs. However, because we were not able to review important details about the work—sample sizes, who was selected to participate, the questions they were asked, and how results were analyzed—and key findings resulting from the work (e.g., areas of agreement and disagreement across participants), we cannot comment on the soundness of the job analysis findings. We also were unable to determine how the job analysis results would be used to inform which tests are most appropriate for use as screening criteria before and during training, nor could we determine

how minimums on those tests would be established or whether bias on the test would be examined.

Although we had few details on the methods OPM would apply, we note that OPM has a long history of work in the areas of job analysis and validation of selection practices. It also serves as a public resource to help in designing valid and unbiased selection practices. For example, the OPM website states the following about physical ability testing:

> Many factors must be taken into consideration when using physical ability tests. First, employment selection based on physical abilities can be litigious. Legal challenges have arisen over the years because physical ability tests, especially those involving strength and endurance, tend to screen out a disproportionate number of women and some ethnic minorities. Therefore, it is crucial to have validity evidence justifying the job-relatedness of physical ability measures. (OPM, undated)

It states the following about job analysis:

> Job analysis is the foundation for all assessment and selection decisions. To identify the best person for the job, it is crucial to fully understand the nature of that job. Job analysis provides a way to develop this understanding by examining the tasks performed in a job, the competencies required to perform those tasks, and the connection between the tasks and competencies. (OPM, undated)

OPM also provides a detailed description in the Delegated Examining Operations Handbook (OPM, 2007) of its job analysis methodology as a public reference on methods for job analysis. That methodology includes a highly detailed and structured questionnaire administered to SMEs (those most knowledgeable about the job) to identify the most critical tasks in a given job and link those tasks to the underlying competencies needed by personnel to be successful in those tasks. The approach described there is generally consistent with recommended practices in job analysis, although again, a close examination of the results would ultimately be necessary to make a final determination on the soundness of the work.

A well-engineered and -executed job analysis lays the foundation for amassing evidence to support a selection system (our recommended Stage 1); however, that alone will not suffice. Additional evidence showing the link between the information collected in the job analysis and the screening criteria is needed (our recommended Stage 3). That link could be established using content validity information if the content validity evidence is strong. However, we could not know the strength of the evidence without seeing the details of the methods used, as well as the results of any data collection efforts and analysis.

For jobs where training serves to screen candidates, it is important to confirm that the training serves as a fair and accurate screening tool. This requires that the standards be applied from one person or one training class to the next. It is possible that training in one class is harder than training in the next (e.g., because of weather, differences in terrain, and differences in simulated mission sets) even if the standards (such as time to completion) are the same. Given this, even if the content appears to be relevant, the minimum standards for performance in the training may not be effective in determining who can meet the requirements on the job. If the training activities are content valid and highly standardized, the scores have equivalent meanings across individuals and classes, and success in those activities is not dependent on irrelevant factors (such as chance events or the performance of one's teammates), then the training can be used for such screening. Evidence to support this should be part of the content validation process, both to inform Stage 3 of our recommended process (validating the selection criteria) and Stage 4 (establishing minimum standards).

The OPM job analysis as it was described to us did not include any plans to consider alternative screening methods beyond those already in place. As a result, we cannot say how well our recommended Stage 2 is being addressed by USASOC's approach. It is possible that activities more closely aligned with the physical requirements of the job and/or gaps in the training activities could be identified through the job analysis. If so, changes should be made to the training. Additionally, the work described to us did not include reviewing the existing minimums on the screening criteria embedded in the training programs (Stage 4).

Lastly, the effort did not include validation of the screening process used to determine who is selected for training. The OPM job analysis findings should be useful in informing decisions about the pretraining screening process, as well as the training screening processes.

CHAPTER SIX

Marine Corps Combat Arms

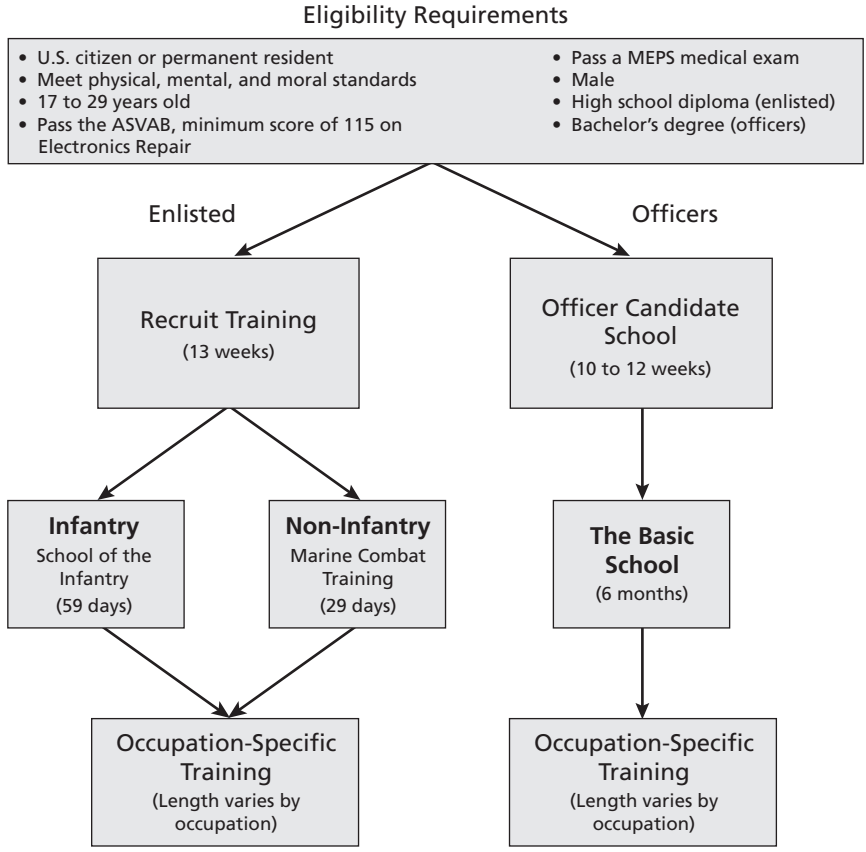
As of May 2013, 32 Marine Corps officer and enlisted primary occupations were closed to women. These occupations accounted for approximately 35,000 active-duty positions and 11,000 reserve positions at that time—roughly 70 percent of them in infantry jobs. In addition, 16 nonprimary occupations were closed. More recently, the group intelligence officer specialty, with a total of about 130 positions, was opened to women.

Occupational Assignment and Screening in the Marine Corps

Figure 6.1 shows the typical path of entry for the combat occupations closed to women. Eligibility requirements include holding U.S. citizenship or permanent residency, being ages 17 to 29, exhibiting good physical condition and moral standing, graduating from high school or holding an equivalent certification, and scoring above a stated ASVAB threshold. Officers must also have a bachelor's degree.

Interested men enter the closed enlisted occupations through their local MEPS. They contract for an occupational career field when they join and receive their specific occupation assignment during recruit training. Recruits choose from the career fields taking new recruits at the time they will enter and for which they qualify based on their aptitude test scores. No occupation-specific physical qualifications exist for entry-level occupations. However, to demonstrate their "good physical condition," before enrolling in Recruit Training, all enlisted recruits

Figure 6.1
Eligibility and Training Path for Marine Corps Combat Arms Branches



SOURCE: U.S. Marine Corps, undated.
RAND *RR1340/2-6.1*

must pass an Initial Strength Test (IST) by completing two pull-ups (men) or a 12-second flexed arm hang (women), 44 sit-ups in two minutes, and a 1½-mile run in 13:30 minutes (men) or 15 minutes (women). However, Marine Corps recruiters recommend to recruits that they strive for well beyond these minimum requirements. No other physical screening is conducted to qualify recruits for entry-level occupations.

After basic training, all enlisted candidates who fulfill the eligibility requirements proceed to Recruit Training (12 weeks) followed

by either the School of Infantry (59 weeks) for infantry occupations or Marine Combat Training (29 weeks) for noninfantry occupations, an abbreviated infantry-training course that qualifies them as "Marine riflemen." After graduating from one of these two training courses, enlisted candidates enter occupation-specific training courses that last for varied time periods depending on the occupation.

Individuals enter officer occupations through four-year colleges, the U.S. Naval Academy, or by transitioning from enlisted to officer ranks. Entering personnel are assigned an occupation no later than basic training. Officer candidates enter the ten- to 12-week Officer Candidate School, followed by The Basic School (six months), and then enter occupation-specific training courses, which, like the enlisted occupation-specific training courses, also last for varied periods depending on the occupation. Officers enter training knowing whether they will be in an aviation or ground combat field, but specific occupational assignments within the field are made roughly halfway through The Basic School.

Given the existing training and occupation assignment procedures, screening to determine which individuals (officer or enlisted) meet occupation-specific physical requirements must be done prior to or soon after beginning basic training. The Marine Corps' plan for integrating women into newly opened positions, submitted to the Secretary of Defense in May 2013, indicated that "the timing and location for administering a screening test for entry into physically demanding occupations now closed to women is dependent upon information learned during the development of the test itself and could range from preaccessions (recruiter-administered) screening to physical screening conducted during recruit training and prior to MOS school assignment" (Mabus, 2013a). The plan anticipated that officer screening would occur during training at The Basic School and enlisted screening would be performed by recruiters, if possible, for those interested in a physically demanding occupational field. If screening by recruiters proved difficult to implement, it would be done during initial recruit training. Even if recruiters perform the initial screening, recruits might be screened again when they reach basic training to confirm their eligibility for physically demanding occupations.

Marine Corps' Process for Establishing Standards for Combat Arms

The Marine Corps established several related but independent efforts to prepare for the integration of women in closed occupations.

1. **Criterion-related validity of PFT and CFT for predicting simulated combat tasks.** This analysis identified the individual-level physical tasks required of individuals in each occupation and the performance standards for successful completion of these tasks. This information was then used to design a study that correlated individual physical screening tests from the PFT and CFT with performance in proxies of the most physically demanding tasks identified.
2. **Opening the infantry training courses for officers and enlisted personnel to women volunteers.** As of the beginning of June 2014, 15 female officers and 199 female enlisted Marines had volunteered for infantry training. Women who successfully complete the infantry-training course cannot be assigned to positions in these closed occupations, so they must pursue one of the occupations open to women. The women who volunteer train in female-only units following the standard training curriculum. This effort includes a survey of participants to assess their reasons for volunteering, attitudes prior to training, and experiences and attitudes when they drop out or complete training. The survey information, along with the performance information routinely collected during training, will be analyzed to assess training attrition or completion and training performance. Although the data and analysis are not sufficient to establish physical standards and screening procedures, the information collected from this effort will supplement the more extensive information from the other two efforts.
3. **Creation of an "integrated task force" to evaluate the performance of gender-integrated ground combat teams.** The Ground Combat Element Integrated Task Force (GCEITF) resembles a battalion landing team and consists of 376 Marine

volunteers, including 77 women, who were assigned across occupations and units in the task force. Each participant has completed training in the occupation for the position he or she will fill on the task force. Participants who had not already completed occupational training did so during summer and fall 2014, and the task force was stood up in November 2014. Following a 20-week unit-training period, individual units consisting of a random sample of individuals and a varying gender mix were evaluated as they rotated through a series of simulated events. The evaluation compared the ability of units with no women versus one or two women to meet performance standards in the events. The performance of individual participants was also measured, compared across units with differing gender mixes, and correlated with individual physical characteristics. Other outcomes evaluated included attrition and injury rates; medical readiness and deployability; and cohesion, morale, and discipline. The data collection plan for the experiment included information that could be used along with the results of the first effort to finalize screening procedures and criteria for closed occupations and assigning Marines in open occupations to ground combat units. When data collection for this project ended, the integrated task force had not yet started its data collection efforts.

In the remainder of this chapter, we describe the methods used in the first and third efforts.

Criterion Validation Study of the PFT and CFT

The Marine Corps assigned responsibility for carrying out this study to Training and Education Command (TECOM), which completed the effort in December 2013. The overall structure of the study, documented in an NHRC and U.S. Marine Corps TECOM report (undated), broadly followed our Stages 1 through 3. However, it relied on existing job descriptions and sought to validate the existing PFT and CFT screening tests rather than identify the most appropriate from a broader set of candidate tests.

Identify Physical Demands

TECOM relied on existing training and readiness (T&R) manuals and programs of instruction (POIs) to identify and describe the physical tasks required for these occupations. In 1995, the Marine Corps established the T&R manuals; in 2011, it developed the Ground Training and Readiness Manual Group (TRMG) Charter Terms of Reference and interim rules to guide regular review, revision, and updating of the T&R manuals and POIs for all ground occupational fields (U.S. Marine Corps, 2011). The validity of the physical task descriptions used as the basis for the Marine Corps' occupational standards development process rests on these procedures for maintaining the T&R manuals and POIs.

The schedule for reviewing T&R manuals and POIs is established annually. Prior to the start of a review, the advocate for the occupation is responsible for reviewing the Mission Essential Task List (METL), which names the essential tasks, conditions, and performance standards required to ensure successful mission accomplishment. A critical step in the review process is the front end analysis, which is initiated on a regularly scheduled basis[1] or sooner if new equipment, organizational or doctrinal changes, evidence of training deficiencies, or other considerations indicate review is needed. During the front end analysis, experts from the occupations review and update the occupation-specific task lists. They also consider whether the same task involves important differences across occupations, what differences may arise in a deployed versus nondeployed environment, and whether there are any equipment changes that affect the tasks.

Additional data are collected on the resulting task list using a survey administered to a sample of job incumbents. Respondents report the time spent performing the task relative to others. Experienced enlisted and officer respondents also report the training required to learn the task in terms of the relative emphasis the task should get during formal training. The survey results are used to determine which are core tasks that should be included in the list of essential skills

[1] TECOM staff indicated during our meetings that the regular schedule is every three years.

required to qualify for the occupation. However, in discussions with the Army and Marine Corps, we heard that ground combat activity in Iraq and Afghanistan rarely involved some tasks that would be more frequent in other types of warfare. Application of the survey results takes this into account. Also, if issues arise during the front end analysis requiring more input than the survey provides, focus groups may be conducted to further explore the issues.

Once the front end analysis is complete, the revised T&R manual and POI are reviewed and approved in a conference involving representatives of the advocate for the occupation, related occupations, the operating forces, and the training centers, as well as SMEs chosen by the occupation advocate. The resulting T&R manual specifies the individual training standards required for collective unit events that in turn ensure performance standards are met in mission essential tasks. The POIs describe in detail the training courses to meet the individual training standards.

To begin the study to develop gender-neutral occupational physical standards, TECOM analysts reviewed the most current T&R manuals and POIs for the primary occupations closed to women. SMEs from each occupation assisted the TECOM analysts. The review had several purposes: (1) to identify physically demanding tasks, (2) to ensure that the description of those tasks was accurate—i.e., to determine whether there were circumstances indicating the documents might require review, and (3) to add any specific information needed about the physical requirements associated with the tasks, such as the weights and dimensions of objects to be carried or lifted. At the same time, the TECOM analysts also conducted a preliminary review of requirements in open occupations.

Of the primary occupations that were closed to women when the study began, ten entry-level enlisted occupations were included in the study. Table 6.1 lists the ten occupations by occupational field. Across these occupations, TECOM identified 32 physically demanding tasks. After the initial review, the Marine Corps opened seven previously closed combat-related enlisted occupations to women in July 2014.

Table 6.1
Marine Corps Ground Combat Enlisted Occupations with Physically Demanding Tasks and Closed to Women

Occupational Field	MOS Code and Description
Infantry	0311 Rifleman 0331 Machine Gunner 0341 Mortarman 0351 Assaultman 0352 Anti-Tank Missileman
Artillery	0811 Field Artillery Cannoneer 0812 Field Artillery Nuclear Projectileman (0811 with nuclear training)
Tank and Assault Amphibious Vehicle	1812 M1A1 Tank Crewman 1833 Assault Amphibious Vehicle (AAV) Crewman

SOURCE: TECOM, Marine Air-Ground Task Force Training and Education Standards Division, 2013.

Validate the PFT and CFT

TECOM opted to use the existing Marine Corps fitness tests (the PFT and CFT) as the predictor tests in its first data collection effort. The PFT has been the Marine Corps' general fitness test for more than 30 years. Responding to a high rate of noncombat injuries in Iraq, the CFT was developed in 2008 to supplement the PFT with a test that is more combat related and improve the physical conditioning of Marines for carrying heavy combat equipment loads. For their use in regular fitness testing for all Marines, both tests are scored using age and gender norming to assess overall fitness. However, for use in setting gender-neutral physical standards for selection into specific occupations, the un-normed test results were used.[2] The components of the two tests are

[2] See Chapter One (Introduction) in Volume 1 of this report (Hardison, Hosek, and Bird, 2018) for a discussion of appropriate use of normed versus un-normed fitness test scores.

- PFT
  - pull-ups—as many as possible (required of men, optional for women); *or* flexed arm hang—as long as possible (required of women as an alternative to pull-ups)
  - crunches—number completed in two minutes
  - 3-mile run—time to finish

- CFT
  - movement to contact event—time to finish 880-yard run
  - ammunition lift event—number completed
  - maneuver under fire event—time to complete 300-yard course incorporating sprints, crawling, carrying casualties, carrying ammunition cans, and throwing a grenade for accuracy; time adjusted for grenade accuracy.

The decision to use these existing fitness tests was motivated by a 2012 study to correlate scores on both tests with performance in ground combat events that all Marines should be able to perform. In that study, TECOM used inputs from SMEs and incumbents in ground combat units to identify three ground combat events to be predicted by the fitness tests:

- MK 19 heavy machine gun lift (72-pound replica)—up to two attempts to lift overhead
- casualty evacuation (165-pound mannequin with 43-pound load)—timed
- 20-kilometer march under 70-pound combat load—completion within five hours.

The heavy machine gun lift and 20-kilometer march were performed while wearing combat gear weighing approximately 70 pounds. The casualty evacuation was performed with a lighter 43-pound combat load.

TECOM tested 2,445 Marines on the three events, including officers and enlisted personnel at the end of boot camp, at the beginning of infantry school, and serving in infantry battalions that had returned

from deployment either four weeks or six months prior to testing. The sample included 424 women. Most of the test subjects were young, 18 to 23 years old. Un-normed individual PFT and CFT test results from the end of basic training (officers) or one month earlier (enlisted) were correlated with performance in the three events, individually and combined, and the analysis also assessed how well a combined PFT-CFT score predicted performance on three ground combat events.

The results revealed significant gender differences in average performance on the heavy machine gun lift and casualty evacuation, whereas almost all men and more than 90 percent of women completed the 20-kilometer march within the five-hour limit.

The report documenting the analysis concluded that the CFT tests were good predictors of performance on the three ground combat events. For men, the PFT three-mile run and PFT pull-ups also predicted the three ground combat events. For women, the PFT run predicted some of the ground combat events, but the flexed arm hang was not a good predictor.

Although past research showed some relationships with the three combat events, TECOM set out to further test the predictive validity of the PFT and CFT for predicting tasks specific to the closed occupations. To do this, TECOM started by categorizing the 32 physically demanding tasks identified for the occupations shown in Table 6.1 according to the type of physical capability required. It next developed a proxy task for each of the identified five task types, as shown in Table 6.2. These were used to simulate the performance that would be required on the job. Although each occupation is associated with different tasks and the level of physical ability required likely varies depending on the task, TECOM opted to treat all occupations as requiring the same job tasks and the same levels of performance on those tasks. That is, it opted to develop a single set of screening criteria for all of the closed occupations. It offered the rationale that any combat arms member might regularly be called upon to perform the physical tasks associated with the other combat arms occupations, not just those required in a specific occupation. Given that the proxy tasks were designed to reflect only the most physically demanding tasks across all of the occupations

Table 6.2
Proxy Tasks for Physically Demanding Tasks in Marine Corps Occupations Closed to Women

Task Group	No. of Job Tasks	Job Task Examples	Proxy Task	Description of Proxy Task
Lift heavy object to above shoulders	9	• Lift M1A1 (Abrams) tank hatches—70 lbs • Assist pushing crewman out of turret from below—115 lbs	Clean-and-press	Lift bar with weights to shoulders and then lift above head • One repetition each to maximum completion at 70 lbs, 80 lbs, 95 lbs, and 115 lbs; participants could elect to skip lower weights • Six repetitions at 65 lbs in 1 minute
Lift heavy object to lower height	19	• Replace track block on M1A1 tank—60 lbs • Lift light armored vehicle strut assembly—three-man team at 135 lbs each	Dead lift	Lift bar with weights to knuckle height • One repetition each to maximum completion at 60 lbs, 70 lbs, 80 lbs, 95 lbs, 115 lbs, and 135 lbs; participants could elect to skip lower weights
Lift and carry heavy object	2	• Lift and carry 155-mm round 50 m in 2 minutes • Lift and carry 100-lbs general mechanics toolbox	155-mm lift/load	Lift and carry 155-mm replica artillery round weighing 95 lbs 50 m, wearing fighting load • One repetition in less than 2 minutes
Lift and load heavy object	1	• Load M1A1 rounds (gunnery skills test, requires five rounds in less than 35 seconds)	120-mm lift/load	Lift 120 mm (55 lbs each) replica tank rounds off 20-in. box, flip, and stack on second 20-in. box • five repetitions in less than 35 seconds

Table 6.2—Continued

Task Group	No. of Job Tasks	Job Task Examples	Proxy Task	Description of Proxy Task
Lower-level entry	1	Negotiate obstacle course wall (training simulation of a lower-level building entry)	Negotiate course wall	Climb over 7-ft wall using 20-in. assist box, wearing fighting load • one repetition

SOURCE: TECOM, Marine Air-Ground Task Force Training and Education Standards Division, 2013.

in Table 6.2, the final set of proxy tasks may not reflect the demands in any single occupation in that table.

The first two proxy tasks, the clean-and-press and the dead lift, simulate the general lifting motions and weights of the tasks they proxy but do not simulate the actual objects to be lifted or the circumstances in which the task is usually done. The last three proxy tasks more closely simulate the objects involved in the tasks they proxy, and the last two—120-mm lift/load and negotiate course wall—capture an important aspect of the circumstances, as they were performed wearing a 40-pound fighting load (roughly the weight of the body armor, helmet, gun, and ammunition). Each of the five proxy tasks was performed in a single or limited number of repetitions, and therefore does not test the ability to perform sustained, physically demanding work.

TECOM collected performance data on the five proxy tasks for 466 enlisted Marines in Marine Combat Training at the School of Infantry–East at Camp Lejeune in North Carolina; in addition, 230 enlisted personnel were tested at Recruit Depot Parris Island and 94 officers at The Basic School at Quantico, Virginia. A total of 790 active-duty Marines were tested, including 410 men and 380 women with an average age of 22. The test subjects were volunteers, but the participation rate among those invited was very high at 98 percent. The most recent PFT-CFT scores for each participant were used. However, because the flexed arm hang showed poor predictive ability in the 2012 study described earlier and to make the test gender neutral, all 790

participants also completed the pull-up component of the PFT during the testing. Prior to testing, participants received instructions and a demonstration of the correct way to perform each task and were given an opportunity to practice using lighter weights.

The analyses included calculating correlation coefficients between individuals' results on each component of the PFT and CFT and their performance on each dichotomous proxy task and for a composite proxy task score equal to the percentage of tasks completed. In addition, summary statistics for each proxy task were provided by gender.

Results showed the easiest task was the deadlift (also up to 115 pounds); all male participants and 99 percent of the female participants were able to successfully complete the deadlift at 115 pounds (i.e., all men and 99 percent of women passed). All male participants passed the 155-mm lift/load task, the 120-mm lift/load task, and the negotiate course wall task. In contrast, only about 70 percent to 80 percent of the female participants passed those tests. By far the most difficult of the proxy tasks, especially for women, was the clean-and-press, which required upper-body strength to lift weights of up to 115 pounds above the head. Eighty percent of the men but only 9 percent of women were able to complete this task successfully. In other words, the only variance observed for men on the proxy tasks was on the clean-and-press. In contrast, for women, there was variance on four of the five proxy tasks.

It is worth noting that the clean-and-press proxy requirement of 115 pounds is not necessarily representative of all of the tasks for which it was intended. As a result, it may not be the ideal measure of whether or not the selection tests are predictive of success on those tasks. To the extent that the 115-pound requirement applies only to certain occupations, this could limit the generalizability of the results. Of the nine physical job tasks for which the clean-and-press serves as a proxy, three require lifting a weight of 100 to 115 pounds, two involve weights of 60 to 85 pounds, and four involve less than 60 pounds. Just about all men tested could handle weights up to 80 pounds, and more than 90 percent could handle 95 pounds. Almost half of the women tested could lift 80 pounds, and 70 percent could do six repetitions of 65 pounds. Although many women could not pass the 115-pound clean-and-press

requirement, it is clear that more would have been successful if the weight had been more closely aligned with these other job tasks.

The report does not present the results for the composite score by gender, but it is clear that 80 percent of men scored 100 percent, and the remaining 20 percent of men scored 80 percent on this measure. It is more difficult to infer what the composite score results were for women, but it seems plausible that the majority scored 80 percent, and much smaller fractions scored 100 percent or below 80 percent. With this pattern of results, the PFT and CFT scores needed to predict success on the proxy tasks depend critically on the maximum required weight that individuals must be able to lift above the head. The fraction of women who could qualify for assignment to combat occupations will vary significantly depending on this single requirement.

The analysis relied on correlation coefficients to measure the association between PFT and CFT test results and proxy performance task outcomes, leaving out the deadlift task because there was no variation in performance on this task (essentially everyone could do it). As the NHRC/TECOM report notes, the lack of variation on most of the individual proxy performance tasks makes it difficult to measure the correlation with the PFT and CFT scores. The analysis showed that performance on the individual PFT and CFT tests, except for the PFT crunches, was highly correlated with the combined proxy task performance score (correlation coefficients ranging from 0.6 to 0.8). The correlation coefficients were approximately 0.70 for the PFT pull-ups and all of the CFT events, 0.58 for the 3-mile run, and 0.37 for crunches.

The NHRC/TECOM report concludes that the CFT events overall are better predictors of performance on the proxy tasks than the PFT events and that the three CFT tests show approximately equal predictive power. Among PFT events, pull-ups predict proxy task performance better than the run, while the run predicts better than crunches. The report concludes the analysis has shown that the PFT and CFT provide a valid basis for determining individual capability to perform physically demanding tasks in closed Marine Corps combat occupations.

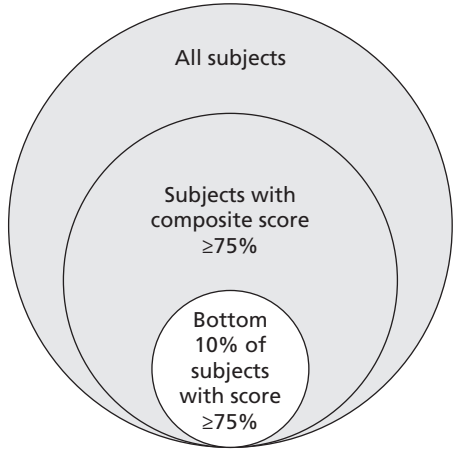
Based on the bivariate correlations, the researchers recommend that applicants for combat arms occupations be screened using an

enhanced IST (which they call the IST+) consisting of pull-ups, the 800-yard sprint, and the ammunition lift. These tests had correlations with the combined proxy task score that were higher than correlations for the crunches test, indicating that they would be better predictors of combined task performance. Their correlations are generally similar to the correlations for the 3-mile run and timed shuttle run, indicating approximately equal predictive power, but the run tests would be more difficult to implement in a wide range of settings, including recruiter stations. Therefore, these data suggest that focusing on the more easily implemented screening tests would have little effect on predictive validity. But again, the regressions were based on the pooled male and female data.

Identify PFT and CFT Screening Tests and Set Minimum Scores for Selection

The NHRC-TECOM report recommends that a version of the existing IST be used to screen Marine recruits interested in the infantry field.

Figure 6.2
Study Population Used to Calculate Cutoff Scores for Infantry Qualifying Test Proposed in NHRC-TECOM Report



SOURCE: Naval Health Research Center and U.S. Marine Corps Training and Education Command, undated.

RAND *RR1340/2-6.2*

The test would include the PFT pull-ups event and the CFT movement to contact (880-yard sprint) and ammunition lift events. The report calculates cutoff scores that might be used to qualify recruits for the infantry field based on the scores for the three PFT-CFT screening events posted by test subjects able to perform the proxy tasks. Specifically, the cutoff score for each screening event was set at the mean score for the lowest decile among only those Marines who performed well on the proxy tests. The researchers defined performing well in two ways and separate cutoff scores were calculated based on each definition. First, they defined it as completing at least 75 percent of the proxy tests (75 percent or higher composite score), as illustrated in Figure 6.2.

Among the group of test subjects with a composite score of 75 or higher, the lowest decile of performers averaged only a single pull-up. The researchers noted that this result reflected the predominance of women in the bottom decile of this group. Therefore, the researchers recommended that this cutoff score be increased to three pull-ups, the existing minimum required of all male Marines on the PFT-CFT test. In contrast, the rest of the predictor cut scores that resulted using the procedure described earlier, represented a more stringent standard than is required on the three PFT and CFT events. The researchers recommended using these cut scores for the remainder of the predictors.

The second way the researchers defined performing well was much more stringent. They defined it as successfully performing all proxy tasks (100-percent composite score). As a result, the cutoff scores based on this smaller subset of test subjects were well above the minimum standard for the PFT and CFT. Of note, the bottom decile in this smaller group averaged five pull-ups. Note also that the gender makeup of the two composite score groups (100 versus 75) were notably different. One-third of those receiving a 75 were women, whereas only 8 percent of those who received a 100 were women. Not surprisingly, the cutoff scores set using this more stringent definition of successful performance would have a greater impact on women than on men.

Our Evaluation of the Criterion-Related Validation Study

The procedures outlined in the terms of reference for the Marine Corps TRMG generally correspond to the procedures associated with

our Stage 1, conducting a job analysis. However, because the TRMG focuses broadly on occupational requirements, the specific information it provides may be incomplete for the purposes of setting physical standards for entry into the occupations. TECOM analysts recognized this and augmented parts of the TRMG by collaborating with occupational experts to identify physically demanding tasks and add relevant information to the task descriptions (e.g., equipment weight, required completion time).

With respect to our Stage 2, the Marine Corps effort did not fully address it. The decision to rely on PFT and CFT scores to predict job performance was made at the beginning of the standards development work, not after the physical job tasks were identified. The 2012 study cited in supporting this decision evaluated how well the six PFT and CFT tests predict performance in three tasks considered relevant to all Marines in the ground combat element. These tasks do not correspond to any of the 32 job tasks for which performance was measured using the proxy tasks, although the machine gun lift used in the 2012 study resembles the job tasks proxied by the clean-and-press test.

With respect to our recommended Stage 3 (validated and select texts), the adequacy of the evidence to support validity rests on the adequacy of the performance simulations used, the design of data collection, and the methods to analyze the data. We discuss each of these in turn.

First, the study used a simulation-based criterion-related validation approach where proxy tasks served to simulate the physical job tasks. In cases where simulations are used as outcomes in a criterion-related validation study, the researchers need to be able to demonstrate links between the simulations and the actual tasks required on the job. In this case, questions could be raised about how well the proxy tasks represented the on-the-job tasks.

Three of the proxy tasks—lift and carry, lift and load, and lower-level entry—were intended to simulate only one or two job tasks. They resemble the job tasks they proxy in that they use simulated equipment, but they do not simulate the typical working conditions. Therefore, these tests have less fidelity than they would if they replicated the actual equipment and working conditions. The other two tasks—

clean-and-press and dead lift—proxy a larger number of different job tasks; therefore, they less closely resemble the actual job tasks. They replicate some of the weights of the equipment used in the job tasks, but not the dimensions or other characteristics of the equipment. They also do not replicate typical working conditions. To the extent that performance is affected by equipment characteristics and working conditions, the proxy tasks could be considered "construct deficient," i.e., they did not capture some important elements of the identified physical job tasks.

The proxy performance tasks required a limited number of repetitions of physically demanding tasks by rested and unburdened subjects in a controlled environment. In contrast, the jobs require that the heavy work be sustained over a longer period of time and performed while rest deprived, with gear on, and in a variety of potentially challenging environments. This must be taken into account in deriving occupational qualification standards from the task performance results. TECOM recognized that additional data collection and analysis would be necessary to track how well the occupational standards predict training and job performance and identify any adjustments in the standards that such analysis may suggest. Longitudinal data with information on the same individual Marines at different points in time would provide a more-definitive validation of the standards than was possible in the development process.

It is also worth noting that the simulations were intended to apply to all closed occupations regardless of whether a given occupation required the task simulated, and no adjustments for differences in task difficulty across jobs were made. For example, the clean-and-press and deadlift proxy tasks both require that participants lift a series of weights up to the maximum weight across all the job tasks for which they proxy. The clean-and-press test proxies for nine tasks in four occupations, with weights ranging from 50 to 115 pounds. In the first round of data collection to validate standards based on the PFT and CFT, participants were not considered to have passed the clean-and-press task unless they lifted the top weight, whether or not their occupation had a job task requiring this weight. Given that the standards established based on analysis of these data will not be occupation-specific

(instead, they will apply to the closed occupations as a group), validity of the resulting standards will depend, at least in part, on the rationale for requiring all Marines in these closed occupations to be able to perform the physically demanding job tasks in all the closed occupations. Otherwise, TECOM should correlate PFT and CFT scores with outcomes for the clean-and-press and deadlift proxy tasks at the relevant maximum weight for each occupation and set occupation-specific selection standards.

Second, with respect to the design of the data collection processes, the sample and the timing of the predictor and outcome data collection matter. The TECOM study collected concurrent validity evidence, meaning that the scores from the predictor tests (PFT and CFT) and job performance measures (proxy task results) were collected at about the same time for the same individuals (the PFT and CFT scores were the most recent scores from their annual fitness tests). The test subjects were trainees instead of job incumbents, and the vast majority were enlistees. As a result, their proxy task performance was probably lower than the performance of incumbents would be and higher than new recruit performance would be. Given this, the timing of data collection in this study is well suited for setting minimum scores for qualifying individuals for ground combat occupations during basic training, when occupational assignments are made. However, for screening at the recruiting station (or MEPS), the minimum scores should be set lower to allow for improvement in physical capability during basic training. If the screening is also done to requalify job incumbents, the minimum scores should be set higher in recognition of physical skills gained on the job.

The male study subjects may have had some advantage in the most difficult proxy task, the clean-and-press. First, male Marines must do pull-ups as part of the PFT and can be expected to include pull-ups in their regular workouts. At the time of the study, the PFT substituted a flex-arm hang for pull-ups for women, and the data reflect the irrelevance of the flex-arm hang for upper-body strength. On net, it seems likely that the pass rates on the proxy tasks for women would have been higher had they trained consistently for tasks requiring upper-body strength. In the future, if pull-ups are a basic requirement for

women to be allowed into infantry and other ground combat occupations, their ability to carry out the job tasks proxied by the clean-and-press test should improve.[3] Second, study subjects were allowed to skip the lower weights for the clean-and-press and dead lift tests if they thought they could easily accomplish those weights and wanted to start at a higher weight.[4] It is likely that most who chose this option were men. In theory, this could affect their performance in lifting heavier weights relative to what it would have been had they lifted all possible weights, because of fatigue effects. In this case, individuals who elected not to skip lower weights (who are likely to be less strong and/or experienced in weightlifting) would find it more difficult to lift the heavier weights because of fatigue alone than if they had not made this choice. In practice, because a high fraction of men could perform all lifting tasks, the option may have had little effect on the results for men. To the extent that women were more likely to begin at lower weights than men, the performance of women would be lowered relative to the performance of men.

Third, the way in which the validation study data are analyzed matters. The validity analysis relied on correlations between scores on the individual PFT and CFT tests and either each proxy task result (pass or fail) or the combined result on all proxy tasks (percentage of tests passed). Unfortunately, the analysis is severely limited by the nature of the proxy task results. The only variation in task performance for men was on the clean-and-press test, which only 20 percent of men failed to perform at the highest weight. Women recorded almost all of the proxy task failures. Because women also on average achieve lower (un-normed) PFT and CFT scores, it is inevitable that analysis of the pooled male and female data will show a positive correlation between the PFT and CFT scores and completing the proxy task. The researchers did not explore the correlation within gender (as is standard and recommended practice when evaluating criterion-related validity). This

[3] Note that expected improvement in these skills as a result of training should be considered in determining the appropriate screening minimums, particularly when the people typically show marked skill improvements with minimal training time and effort.

[4] This option was offered to everyone.

is unfortunate because gender is likely at least as highly correlated with task performance in these data as the PFT and CFT test results. Thus, we cannot rule out the possibility that the correlations the results report in support of the validity of the PFT and CFT are simply an artifact caused by the gender differences in the task performance data. That is, the positive relationships might not hold within gender. However, those relationships should hold within gender for a test to be considered valid and fair for both groups. Even if the researchers had explored within-gender correlations as we recommend, given that there is zero variance in male performance on nearly all proxy tasks, no correlations for men would be able to be calculated (i.e., the correlation would be zero). This inability to explore relationships for men at all is a major limitation of the study results, one that raises questions about whether the screening minimums would be even considered valid for the male population.

An analysis of the data for women only would provide information on the validity of the PFT and CFT battery test for determining which women might be capable of performing physically demanding tasks in closed occupations. However, this would not support the validity of the PFT and CFT for setting gender-neutral physical standards. As we discuss in our overview of methods for developing occupational physical standards (Hardison, Hosek, and Bird, 2018), instituting valid gender-neutral standards means not only employing the same tests and selection criteria for men and women but also ensuring that the tests and criteria are equally effective in predicting which men and which women will be able to perform the required physical tasks—i.e., that they are gender unbiased. To determine whether the PFT and CFT would be valid and unbiased when used for occupational screening, additional testing is required using methods that can differentiate the performance of men and women. Such testing should include male subjects from the same occupations as the female subjects and physical task performance measures that can distinguish different physical capabilities among men.

The NHRC/TECOM report concludes that, with a few exceptions, the PFT and CFT scores are good-to-excellent predictors of job task performance. We question this conclusion for two reasons. First, as

we already described, the gender pattern in the data raises issues in interpreting the correlations as measuring the ability of the PFT-CFT test to predict which individuals of either gender will be capable of job task performance. Second, the standard they cited for considering a correlation to be a valid predictor—a correlation coefficient of 0.30 to 0.40—is not applicable in this context. The source they referenced for the standard was a RAND report (Hardison, Sims, and Wong, 2010) that discusses the Air Force Officer Qualification Test, a cognitive aptitude test. Correlation coefficients vary widely depending on the performance outcomes and types of screening tests involved and a number of other factors. Correlations between cognitive aptitude and on-the-job performance can be around 0.30 to 0.40. In contrast, tests of physical aptitudes, when correlated with physical simulation activities in a laboratory setting, can show correlations ranging from 0.60s to 0.90s, depending on the complexity of the simulations, the timing of the predictor and outcome measures, and many other factors. With all of these other factors held constant, correlation coefficients can be useful for comparing the relative predictive ability of alternative screening tests, but correlations should not be expected to be comparable across tests of different individual capabilities or even different types of validation study designs.

Performance on the three tests proposed for the IST is highly correlated (measured across all test subjects, the pairwise correlations are all about 0.75). In cases such as this, regression equations can be useful in teasing apart which tests are the best to include to maximize prediction, and which tests are redundant. In fact, research has shown that physical tests can overlap significantly in their ability to predict simulated task performance—i.e., some tests can capture the same predictive power as other tests (Vickers, Hodgdon, and Beckett, 2008). In cases where there is such overlap, such factors as ease, cost of implementation, and adverse impact should also guide the choice of which tests to use. NHRC did not employ regression analysis to explore the overlap in the predictive power of the correlated PFT and CFT tests.

With respect to our recommended Stage 4 (establishing valid minimum qualifying scores), the recommended approach was simple to understand, although because of some of the limitations described earlier, it could be challenged. Minimum scores to qualify for an occupation

should be set at the minimum that corresponds to acceptable on-the-job performance. The cutoff scores calculated in the NHRC/TECOM report reflect this criterion so long as successful accomplishment of the proxy job tasks captures the ability to perform on the job and the data collection and analysis methods are appropriate. To the extent that the proxy tasks were easier or more difficult to perform than actual job tasks under expected working conditions, the cutoff scores would be set too low or too high. Too-low scores would qualify individuals who fail to pass training or perform poorly on the job. Monitoring training and job performance over time should quickly identify any need to raise the cutoff scores. Too-high scores, which would keep out people who could have succeeded, would be more difficult to detect.

The NHRC/TECOM report does not address the accuracy with which the suggested screening test identifies who can perform job tasks and who cannot, given the calculated cutoff scores. Any screening test yields false positives (people who qualify on the test but cannot meet the performance standard) and false negatives (people who do not qualify but would be able to meet the standard). Higher minimum qualifying scores yield fewer false positives and more false negatives. Those qualified for the occupation have higher success rates, but individuals who would be able to do the job are denied the opportunity. To gain some insight into what trade-off is anticipated with the occupational physical standards adopted by the services, it would be useful to compute the percentage of test subjects who would be false positives and false negatives given their screening test and proxy task performance results.

Lastly, the study put forth recommended screening minimums for assignment to the closed ground combat jobs, but those minimums were not specific to the occupation or career field. The study did not provide evidence similar to that used in a job analysis to show how often Marines in combat occupations might perform the physically demanding tasks of other ground combat occupations or how critical their ability to do so would be.

The GCEITF Study

The Marine Corps did not plan to make a final decision on a physical screening test (PST) until information from the GCEITF would be

available, toward the end of FY 2015. The GCEITF study was in progress when we completed data collection for this report, so we did not know the results or how they would inform the final screening process for selection into combat occupations. We did, however, have detailed information about the design of the experiment and the analysis plan, so we could evaluate how the results will relate to setting standards for selection into the Marine Corps' closed occupations.

The entire study, which was carried out after the unrelated TECOM study, was essentially a criterion-related validation study using simulated real-world performance as the outcome to be predicted. Thus, the entire description of the study's methodology following relates primarily to our Stage 3 (i.e., validation of the predictor tests). Preparation for the study included elements of Stages 1 through 3, for which we had less complete documentation of methods.

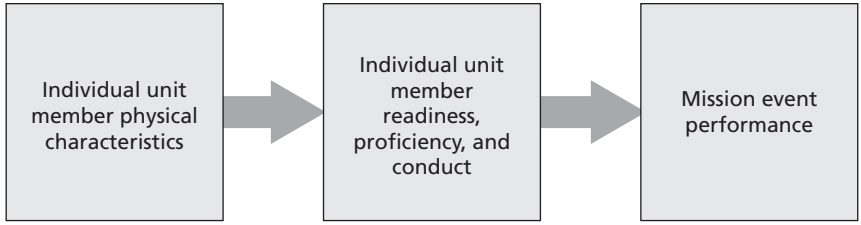
Study Objectives

Detailed objectives of the GCEITF were stated in the research protocol for the study prepared by the Marine Corps Operational Test and Evaluation Activity (MCOTEA), which oversaw the experiment with assistance from researchers from the Center for Naval Analyses, the NHRC, and academic consultants.

This study had two broad objectives, and its design was not specifically targeted to establishing physical standards. The first objective was to determine whether units assigned some women perform any differently than all-male units. Note that all participants had completed occupational training, so this objective did not include determining which men or women can complete training and perform in a unit. The second objective focused on individual abilities and both individual and unit performance, so the evidence collected toward this objective was relevant for establishing and validating gender-neutral standards. The second objective is the one more directly relevant to the purpose of our research (i.e., establishing standards), and so we will focus our discussion on the research plans relevant to this objective.

The study was designed to explore the relationships shown in Figure 6.3—individual physical characteristics impact mission outcomes through individual readiness, proficiency, and conduct. Thus,

Figure 6.3
Hypothesized Relationship Between Individual and Unit Attributes in the Ground Combat Element Integrated Task Force Study



SOURCE: MCOTEA, 2014.
RAND *RR1340/2-6.3*

a physical characteristic like upper-body strength affects how well the individual can perform physically demanding mission tasks and the likelihood of an injury rendering the individual not ready. Less obvious is the hypothesized relationship between physical characteristics and conduct. This is a broad measure of other performance attributes including attitude and leadership. The premise is that someone struggling to carry out assigned tasks and affected by extreme fatigue may not perform as well in these other dimensions as that person otherwise would.

Select Study Participants

The experiment was designed to represent a ground combat element component, with a rifle company (rifle, machine gun, mortar, and assault weapon launcher squads) and other units associated with a battalion landing team (artillery, tanks combat engineers, light armored reconnaissance, AAVs). The design called for each type of unit to conduct multiple trials of a series of mission events to determine whether and how outcomes differed for integrated versus male-only units. Each trial involved three units of the same type performing a set list of tasks associated with a particular mission. The matched units included one male-only unit, a unit with one female member, and a unit with two female members. In the integrated units, the ratio of men to women ranged from 11:1 to 1:1, depending on unit size and whether the unit had one or two women assigned.

A random sample of volunteer participants was selected for the units and specific roles within units for each trial. Sampling was by replacement, so an individual participant had the same chance of being selected for each trial, unit type, and role. The only exception was when more than one trial was conducted at the same time. This procedure was designed to ensure that unit leadership and the individuals assigned to the units did not affect the overall comparison of performance by unit type (men-only, one woman, two women). The trials were conducted at the same times and under the same conditions for the different unit types.

The numbers of trials per mission event and participants in each event were based on a power calculation using information derived from operational tests (if available); otherwise, from live demonstrations of the events. The power calculations were designed to detect a 30-percent to 46-percent difference in the measured outcomes for most events, depending on the outcome and how it is measured. These effect sizes were chosen based on the input of Marines with expertise in the tasks required for each mission event. Based on the power calculations, the researchers estimated that 231 male participants and 77 female participants would be required; however, to allow for potential attrition, injury, or other participant nonavailability, the actual sample was somewhat larger.

Participation was voluntary. All Marines serving on active duty who met the following criteria were invited to participate:

- E-5 or below and fewer than nine years of service
- full-duty status
- closed MOS volunteers (men only): capable of achieving the third-class PFT requirements for men ages 17 to 26; and completed MOS training
- open MOS volunteers: completed MOS training.

Women volunteering for closed MOS experimentation could be new recruits or serving in open occupations prior to the study. They were chosen to be as comparable to male participants as possible, depending on the number of female volunteers. Prior to participat-

ing in the study, the women selected had to complete the MOS training associated with the closed occupation they would fill. The training used the standard course of instruction but was segregated by gender.

One important difference between male and female study participants was on-the-job experience. Male participants could include some Marines just out of training, but many had unit experience and some would have deployment experience. Because the occupations were closed to women, by definition, the female participants lacked unit experience. The three-month unit-training period scheduled before the trials began would have narrowed the experience gap, but not necessarily eliminated it. It is also possible that female Marines with prior experience in an open occupation may have found that experience helped their performance in the experimental unit. Other systematic and unavoidable differences between female and male participants likely included physical characteristics (e.g., scores on the different PFT and CFT elements, height and weight) and perhaps other individual characteristics, such as ASVAB scores.

The study protocol notes that, because participation was voluntary, participants were self-selected. Why participants volunteer for research can affect their motivation, for example, and participants may differ from the population they are drawn from in ability or other characteristics that affect study outcomes. The results of this study will depend on whether volunteers were representative of all personnel in the ground combat occupations (including women who would request the occupations in the future). Of particular concern is whether self-selection among men versus women differed in ways that would affect study findings.

Design Mission Event Trials and Performance Measures

A total of 50 mission events were chosen. Each event consisted of a series of tasks that included the most difficult physical tasks ground combat personnel must be able to perform. Table 6.3 lists the events by occupation. To select and specify these events, the researchers drew on existing training events, which are described in T&R manuals and reviewed on a regular schedule. Selection of the events involved a number of Marine Corps organizations: Plans, Policy, and Organiza-

Table 6.3
Mission Events by Occupation for GCEITF Study

Occupation (MOS)	Events
Machine Gunner (0331)	• Provide suppressive fires with medium machine gun • Provide suppressive fires with heavy machine gun
Rifleman (0311)	• Conduct ground attack (with MOS 1371) • Conduct defensive operations • Conduct dismounted movement in mountainous terrain
Mortarman (0341)	• Provide indirect fires with 60-mm mortar • Provide indirect fires with 81-mm mortar
Assaultman (A351) Anti-Tank Missileman (352)	• Provide offensive fires with Shoulder-Launched Multipurpose Assault Weapon • Provide offensive fires with TOW missile weapon system and Humvee
Combat Engineer (1371)	• Conduct breaching • Conduct counter mobility operations • Conduct dismounted route sweep operations • Destroy captured arms and ammunition with explosives
Light Armored Reconnaissance Crewman (0313)	• Vehicle recovery and tow operations • Prepare vehicle for combat • Engage main gun targets • Evacuate wounded crewman • Conduct maintenance actions
M1A1 Tank Crewman (1812)	• Reload main gun • Manually manipulate turret and main gun • Prepare commander's weapon station • Conduct crew evacuation • Conduct crew operation • Conduct vehicle recovery • Conduct ammunition resupply • Engage offensive targets • Transfer ammunition • Employ loader's M240 machine gun • Evacuate wounded crewman • Conduct maintenance actions

Table 6.3—Continued

Occupation (MOS)	Events
AAV Crewman (1833)	• Secure AAV for transport with chains • Remove chains and unsecure AAV • Conduct AAV water recovery • Conduct AAV land recovery • Load weapons and ammunition • Conduct immediate and remedial actions on weapon • Conduct simulated reload • Evacuate wounded crewman (two events) • Conduct maintenance actions
Field Artillery Cannoneer (0811)	• Emplace • Prepare ammunition • Fire mission • Position improvement • Redistribute ammunition • Shift out of traverse • Displace howitzer artillery piece • Remove unfired projectile • Evacuate wounded crewman • Conduct maintenance actions

tion; TECOM; Ground Combat Advisory Groups; and the 1st, 2nd, and 3rd Marine Divisions. Although not a full job analysis, this preparatory work is related to our Stage 1.

The researchers embedded these events in schedules intended to simulate the mix and pace of activity each type of unit would experience during actual operations. For example, the rifle squad schedules encompassed 48 hours with an attack on the first day, a 7-kilometer march and holding a defensive position on the second day, nighttime bivouacs both nights, and various nonexperimental activities (not requiring physical exertion) interspersed with the experimental activities. In contrast, the artillery schedule encompassed only four hours but included the full list of mission events. MCOTEA carried out the trials at three locations: Twenty-Nine Palms, California (desert terrain); Camp Pendleton, California (varied coastal terrain); and Bridgeport, California (mountain terrain). The number of trials per event varied from 14 to 40, as required to achieve the desired statistical power. A male-only unit, a unit with one woman, and a unit with two women

carried out each trial. As we described previously, the individuals in each of these units were randomly chosen from the pool of participants for that occupation by gender for each trial.

Event-specific performance outcomes were measured at the unit and/or individual level (depending on the event and outcome), and some individual- and unit-level measures were collected using data spanning multiple events (see Table 6.4). Across trials for each event (and event subtask), the data measured unit performance by gender composition of the unit and individual unit-member performance. The event-specific measures were collected at the unit and/or individual level, as appropriate for each event.

The University of Pittsburgh's Neuromuscular Research Laboratory collected extensive physiological data on GCEITF participants, including measures of flexibility, aerobic capacity, and stress, and they

Table 6.4
Performance Measures for Ground Combat Element Integrated Task Force Study

Event-specific measures	• Elapsed time • Rate of movement • Distance covered • Percentage of quantity accomplished (e.g., rounds fired, targets hit) • Self-reported fatigue following the event (7-item scale)[a] • Self-reported maximum workload (7-item scale)[a]
Measures spanning multiple events	• Individual readiness: percentage of days available for duty • Unit readiness: percentage of days available across all participants • Commander assessment of individual proficiency, derived from unit diaries for marking period • Commander assessment of individual conduct, derived from unit diaries for marking period • Incidence of individual misconduct, derived from unit misconduct reports

SOURCE: MCOTEA, 2014.

[a] No additional information on the scale items were provided to us. These measures were taken from the Crew Status Survey, developed by the Air Force to assess fatigue and workload in flight crews, which has been found to correlate well with other measures (Charlton, 2002).

administered the PFT and CFT and other screening tests. We were not given a description of the physiological data collection, including such important details as when the data were to be collected (e.g., before and after collective training, during events or between events). Scores on earlier PFT and CFT tests from basic training and occupational training also were available in personnel records. The analytic plan called for these data to be used to analyze injury rates and performance in the simulated events, as well as to identify training approaches to increase physical performance and success rates.

Analyze Experimental Data

Earlier in this chapter, we described the two principal purposes of the experiment: (1) to determine how assigning women to ground combat units affects collective unit-level performance in simulated mission events; and (2) to measure the relationships between individual physical characteristics, individual outcome measures, and collective outcomes measures. The second purpose is the most directly relevant to setting individual-level physical standards. However, the experiment was designed principally with the first purpose in mind, and the analytic methods described in the research protocol were more fully developed for that same purpose.

In its research protocol, MCOTEA described several planned analyses to explore the relationship between individual physical characteristics and both individual and unit performance. We were told that this information would be used with the results of the study of gender-neutral physical standards described at the beginning of this chapter when the Marine Corps decided what physical standards to set prior to the January 2016 deadline for opening ground combat occupations to women. The planned analyses included:

- Comparison of the distributions of the individual task performance measures by gender. Most of the individual-level metrics identified in the protocol (e.g., rate of each individual rifleman's movement to firing position and on-target percentage of each A1M1 tank crewmember) measured either elapsed time or rate of movement performing a task or, less frequently, on-target firing

percentage. The other individual measures were collected at the event level, instead of the task level, and consisted of self-reported fatigue and workload level during the event.
- Comparison of the distributions of individual performance and conduct during the overall event, as evaluated by unit commanders during the events. The protocol did not indicate whether the distributions would be compared for all occupations and events combined or by occupation, event, or both.
- Regression analysis to estimate the relationship between gender and both individual and unit readiness, controlling for other variables (examples in the protocol include weather, prior experience in the occupation, the team role the individual is assigned to, and level of participation in the experiment to date).
- Analysis of variance (ANOVA) to determine whether individual task proficiency (i.e., the outcome measures of individual performance and conduct during the overall event) is correlated with collective performance in events. As with the regression analysis, the ANOVA analysis would control for other variables.

Overall, the proposed analytic methods are reasonable, given the experimental design and the stated objectives for measuring individual and unit performance during the experiment. The data recorded multiple individual performance measures for the same individuals in different trials of the same event and across different events. The individual participants were randomly combined with other participants to form units and randomly assigned to different roles in the units. The protocol anticipated that randomization would assure the observed data satisfy the requirement that each trial of each event be independent of the other trials for the same event. This is because there were unlikely to be identical units (the same members assigned to the same roles).

However, there is still reason to question whether the observations were fully independent. If, as seems likely, the same individuals systematically do better or worse across trials and even roles and events, the measures taken for the same individual would have been correlated and the statistical power for the experiment lower than expected. Similarly, persistently higher or lower performance by the same individual

would cause some correlation across unit-level performance measures for all units to which the individual was assigned. The protocol discussed methods that take into account persistent differences across individuals in analyzing individual performance data but not in analyzing unit performance data.

Our Evaluation of the GCEITF Study

The MCOTEA protocol did not describe a clear plan to employ any of the selection test validation methods in our recommended Stage 3. It did not lay out a plan to use the data to determine the relationship between physical tests that could be used to screen recruits and individual performance during the experiment. The protocol did mention that analyses of the relationship between individual physical characteristics and performance outcomes would be conducted, but it did not describe the specific methods that would be used for this purpose, nor did it state how the results would be utilized.

However, the experiment did provide the data necessary for conducting this type of analysis. Regression analyses (similar to those proposed in the planned analyses described in the third and fourth bullets in the previous section) could be conducted to identify the best predictors or a combination of predictors for use in selection. The potential predictors could include PFT and CFT scores (taken at various times), the IST (a subset of the PFT and CFT tests), and other physical aptitude measures collected during or prior to the study. Outcomes could include any of the individual outcome measures collected during the experimental unit (e.g., individual-level performance in a given event, self-reported fatigue and workload, individual readiness, and/or conduct). For example, regression analysis could be used to estimate the relationship between IST scores in BCT and the commander's assessment of individual proficiency. Separate regressions could be run for predicting each performance outcome, or multiple outcomes could be aggregated using a method such as factor analysis to limit the number of regressions and minimize the chance of identifying spurious statistical relationships. As proposed in the planned analyses described in the third and fourth bullets in the previous section, other factors, such as weather, also could be controlled for in the regressions. These regres-

sions could also be explored by gender to determine whether the same physical test scores predict equally well for men and women.

Lastly, care should be taken in interpreting findings showing that women did not perform as well as men in the unit events. All the women could have had satisfactory performance even though their average performance was lower than the average performance of the men. Alternatively, some of the women could have performed at an unsatisfactory level. In either case, the first place to look is the training program. In theory, all of the male and female participants graduated from the same (but segregated) MOS-specific training. Gender-based performance differences in the unit events during the experiment could reflect differences in difficulty of the standards set during the gender-segregated training. It is possible that having a standard course of instruction may not be sufficient to ensure the same performance of training graduates. In addition, if a noticeable proportion of participants (male or female) failed to perform satisfactorily, it would raise questions about why the training did not adequately prepare personnel to do the job. Assuming that the expectations for what constitutes satisfactory performance in the unit events were well justified (e.g., through a systematic process involving consensus among multiple SMEs), this would suggest that the minimum standards for graduation from training should be revisited.

Our Evaluation

The Marine Corps planned to rely on the results of two independent studies in developing physical standards for its closed ground combat occupations. The first study explored the correlation between scores on the PFT and CFT tests and simulated individual physical task performance. This study generally followed the basic stages we identified in our review of standard methods. It led to a set of recommended screening tests and minimum qualifying scores for selection into these occupations. Although the process generally aligned with our stages, we identified several limitations in the data and the analyses that could affect the validity of the suggested standards.

The CGEITF was expected to provide additional data and analysis that would address these limitations. MCOTEA designed the experiment primarily to determine whether assigning women who successfully complete training to ground combat units affects unit performance. However, the data being collected could support analyses other than those described in the research protocol. These analyses have the potential to strengthen and supplement the information resulting from the first study. We note, however, that our assessment is based only on the design and analytical plans for the experiment. Without seeing the actual data, methods, and results we could not fully evaluate how useful the experiment turned out to be for this purpose.

CHAPTER SEVEN

Marine Corps Special Operations Forces

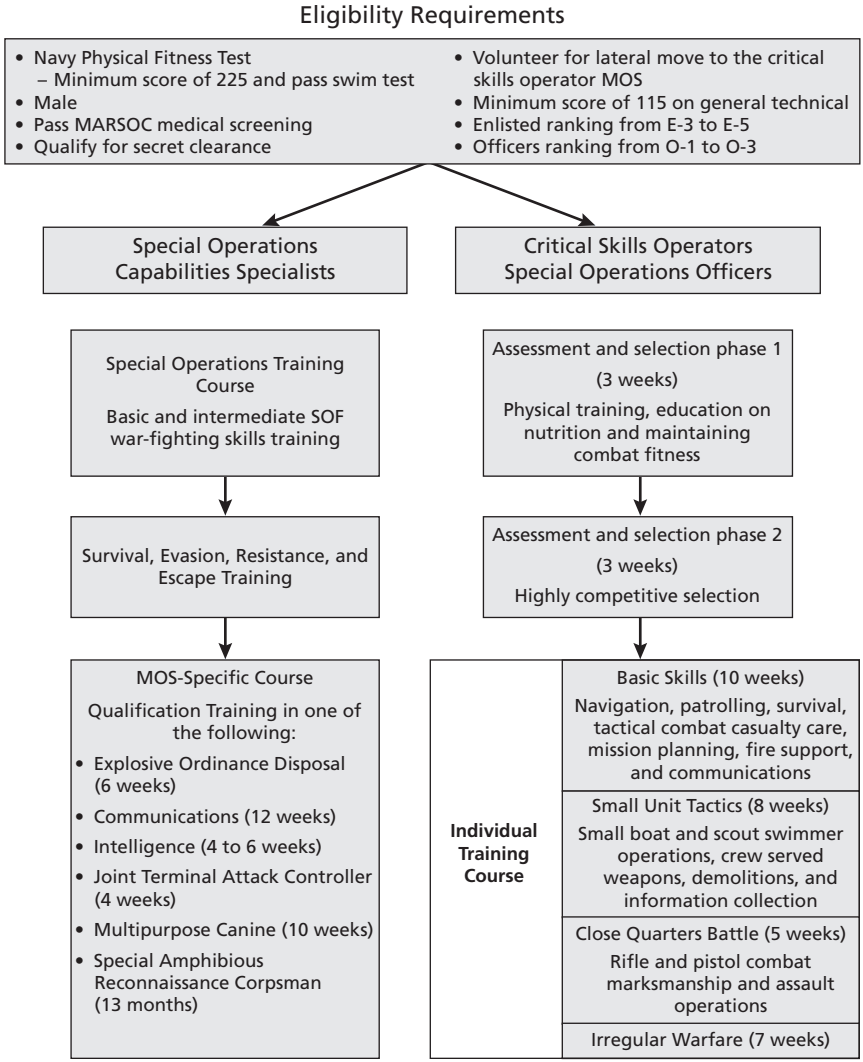
The Marine Corps Forces Special Operations Command (MARSOC) has two closed position types: closed occupations and closed billets. Critical skills operators (CSOs) and special operations officers (SOOs), also known as Raiders, are the only closed MARSOC occupations, and they account for about 900 positions in MARSOC. Special operations capabilities specialists (SOCS) and special operations combat service specialists (SOCS-S) account for the remaining closed positions. They are temporary assignments or billets filled by members of the broader Marine Corps community.

Occupational Assignment and Screening

The CSO/SOO and SOCS selection and training pipelines are displayed in Figure 7.1. Entry-level Marines cannot apply for any of the closed MARSOC occupations. Only those who have served for some time (typically at the rank of E-3 to E-4 or O-3) can apply. As with most of the other special operations forces, selection is carried out through a top-down process.

SOCS personnel are first assigned to MARSOC through award of the 8071 secondary MOS three to 12 months before assignment to MARSOC for an average tour of 39 months (authorized for up to 60 months). They hold one of several specialties (intelligence, communications, explosive ordinance disposal, dog handler, or fire-control specialist) and receive special operations training plus additional specialized training in their specialty prior to serving in the MARSOC billet:

Figure 7.1
Eligibility and Training Path for Marine Corps Special Operations Forces



SOURCE: MARSOC, undated.

RAND *RR1340/2-7.1*

explosive ordinance disposal—six weeks; communications—12 weeks; intelligence—ten weeks; joint terminal attach controller—four weeks; multipurpose canine—ten weeks; special amphibious reconnaissance and independent duty corpsman—13 months. Those who complete the training serve in an extended tour of service with MARSOC. When that tour is complete, they return to the usual assignments in their MOS.

SOCS-S training is shorter than SOCS training. This training is geared toward developing skills in joint and interagency work and operating in the special operations context before entering a MARSOC billet. SOCS-Ss complete a standard-length assignment with MARSOC. Following that tour, they return to other assignments in their MOS.

For CSOs and SOOs, the application and training process is significantly longer and more stringent. Members of any MOS with the requisite minimum years of service can apply for entry into these occupations, and typically about two-thirds of applicants come from non-infantry MOSs. In addition to having the requisite minimum years of service, applicants for these positions must have been deployed at least once, and they must agree to a specific term of service commitment. As shown in Figure 7.1, applicants must also meet a variety of criteria, including ASVAB score, clearance, and medical screening criteria and passing the minimal PFT score of 225 (the mean applicant score is around 270) and the MARSOC swim assessment (300 meters in uniform and treading water for 11 minutes).

Senior MARSOC personnel judge applicants holistically on these entry criteria and only those judged most competitive are selected to begin training. MARSOC has not provided any additional details on how the PFT information is used during the selection process. However, MARSOC's training website advises that candidates should be able to demonstrate a PFT score of 250 or higher and maintain a 4-miles-per-hour (15-minute mile) pace with a 45-pound rucksack regardless of distance prior to entering training (MARSOC, Marine Raider Training Center, undated). Similar to the USASOC occupations, although there are minimum PFT requirements, the majority of the screening for physical ability occurs during the training itself.

The CSO and SOO candidates who are selected participate in the three phases of MARSOC training to assess each candidate in terms of these desired attributes: integrity, effective intelligence, physical ability, adaptability, initiative, determination, dependability, teamwork, interpersonal skills, and stress tolerance.

Phase 1 is a three-week course that includes physical training in running, swimming, and hiking, as well as classroom instruction and hands-on application of Marine Corps, MARSOC, and special operations knowledge. Recruits are screened after Phase 1 for their ability to enter the three-week Phase 2 course: assessment and selection. This phase is a highly competitive evaluation to identify which Marines have what it takes to enter these occupations. MARSOC staff rank recruits based on their performance following the standard Marine method for assessing enlisted personnel for retention and promotion, including fitness reports for officers and pro/con performance marks for enlisted applicants.

Marines who complete Phase 2 are selected to enter Phase 3, the 34-week Individual Training Course (ITC). Phase 3 is a nine-month course in which enlisted personnel (operators) and officers gain special operations knowledge, skills, and strategic awareness. After the completion of the ITC course, enlisted Marines are awarded the 0372 CSO MOS and attend a 26-week basic language course. Officers attend a four-week Team Commanders Course, after which the 0370 SOO MOS is awarded. Following the language course or Team Commanders Course, CSOs/SOOs attend a three-week Basic Airborne Course. They are then part of the operating forces and proceed to advanced specialty skills courses for enlisted or officers.

MARSOC's Process for Establishing Standards for Special Operations Forces

The following are the steps that MARSOC outlined in its plan to address the requirement for valid, gender-neutral standards:

- Conduct detailed job analyses for MARSOC positions of interest (SOOs, CSOs, and SOCSs).[1] This includes identifying critical job-related tasks for each MOS or type of assignment, linking those tasks to specific job duties, identifying critical job-related KSAs, and linking those KSAs to specific tasks.
- Validate the ITC for CSOs/SOOs and Special Operations Training Course standards. This includes identifying minimum entry qualifications for each, using content-based validation to evaluate the validity of the course standards, and developing new course standards and training events.
- Validate the assessment and selection course standards. This includes identifying the selection factors/screening tests, collecting trainee performance data during the assessment and selection and ITCs, and using a hybrid content/criterion-based validation approach to evaluate how well the screening tests predict who can successfully execute the job duties.

Although MARSOC staff members described the validation plans as including a hybrid of content-based validation and criterion validation approaches, they acknowledged that there might not be time to complete any criterion validation work prior to the mandated deadline. In addition, it is worth noting that their planned efforts were solely directed at validating the selection that occurs during the training courses. When our data collection ended, no plans were in place to validate the processes used to screen people prior to entering training.

Like USASOC, MARSOC contracted OPM to execute its validation plan. OPM began the job analysis and standards validation in November 2014 and had scheduled its completion by May 2015. Because of the timing of OPM's contract initiation, we were unable to

[1] MARSOC is also asking OPM to review enablers (personnel in non-SOF MOSs who are assigned to SOF) because they can be assigned at the unit level and must have the physical capability to operate with the unit. MARSOC's current practice is to assign enabler personnel to units based on an informal assessment, but it would like to develop more formal criteria. MARSOC will adopt enabler personnel selection standards based on the OPM work or, if that proves not to be feasible, it will use the regular infantry selection criteria when those are developed.

review its work prior to completing our data collection efforts. However, OPM's statement of work (SOW) was provided for us to review, and it serves as the basis for the description of the work provided in the following section.

Identify Physical Demands

OPM identified several deliverables in the job analysis methodology. The first two deliverables consisted of initial task, competency, and physical ability lists for each occupation. The plan called for the first task list to be created by (1) reviewing relevant existing documentation provided by MARSOC on each occupation or assignment type (e.g., job descriptions, training documents, prior job analyses); (2) reviewing existing scientific literature on physical abilities in general, and OPM's competency lists and other research literatures that might relate to the requirements of the MOS; (3) conducting site visits to observe job incumbents performing job duties, interviewing incumbents and supervisors; and (4) observing the way existing screening tests (if any) are implemented. The results of this process would then be used to generate the first deliverable—a starting list of tasks, competencies, and physical abilities required for success in each MOS or assignment type.

Next, using this initial list as a starting point, the OPM plan was to hold SME panels (likely spanning two days) for job incumbents to review and revise the list. Following that, additional SME panels with supervisors would be convened to verify that the content of the list resulting from the incumbent SME panels is accurate. The list resulting from these SME processes would be the second deliverable provided to MARSOC. It would also serve as the foundation for the next step, the online surveys.

Two online surveys were to be produced and also delivered to MARSOC, one for supervisors and one for incumbents. Both groups would be asked to rate the tasks on importance and the competencies and physical abilities on importance, whether they are required for entry into the occupation, and the need for training in them. Incumbents would also be asked to rate the frequency of the tasks. Other scales might also be included. Analysis of the survey would be geared toward defining the critical competencies, physical abilities, and tasks

for various levels of the MOS or assignment type. The list of critical tasks, physical abilities, and competencies incorporating the survey results would be the fourth deliverable.

The list incorporating survey results would then be reviewed by a new SME panel of incumbents and supervisors to ensure that the list is consistent with their understanding of the job and to resolve any disagreements and inconsistencies in survey responses about what is needed. The SME panel would also be called upon to provide task-competency linkage ratings to establish the relationship between the tasks, the competencies, and the physical abilities. After the SME panel, a final list of the competencies, tasks, and physical abilities, along with the competency linkage findings, would be delivered to MARSOC.

The very last deliverable resulting from the job analysis was a report documenting all of this described work.

Validate Training Tests

As we discussed earlier, MARSOC uses panels of senior personnel to select candidates for entry into training. The work plan for developing and validating standards did not address this initial selection process. Instead, it focused on steps to validate the relationship between performance on tests conducted during the early training stages, which experience a high dropout rate, and ability to perform the required job tasks. The plans provided to us for OPM's validation step were far less specific than the plans for the job analysis. The validation plans acknowledged that both content and criterion-related validation strategies could be used—in addition, that the more evidence collected to support the content and criterion-related validity, the better the support. However, they also noted that the best strategy would depend on the types of tests given during training, and OPM was not told what existing or proposed tests might be considered prior to writing up its statement of work.

If content validation were pursued, it would have included a panel of testing-and-assessment experts reviewing the tests and testing materials (including how they are administered and how scores are assigned and used) and the job analysis findings. The panel also would be asked to provide ratings on how well the tests represent the domain that they

are supposed to measure and their relevance for performance required in the job, based on the job analysis results.

OPM noted that if criterion-related validation were to be pursued, certain data would be necessary: sufficient samples of personnel, appropriate outcome measures (e.g., job performance information), and test score data that meet specific statistical criteria. OPM planned to work with MARSOC to determine whether these requirements could be met. OPM also acknowledged that the existence of these data may differ across MOS/assignment groups, which would necessitate group-specific approaches to validation.

Regardless of the methods chosen, OPM promised to provide a technical report documenting the steps taken to validate the tests.

Set Standards

OPM explained that upon completion of the job analysis, physical performance standards (i.e., the threshold levels of the abilities needed to perform the physical tasks identified in the job analysis) would be set. However, OPM also explained that it does not have personnel with the physiological and medical expertise necessary to set those standards. OPM instead planned to work with specialists designated by MARSOC (such as exercise physiologists) to ensure that any standard setting outcomes would be based on the job analysis. No further information about how the standards would be established or how they would be used to establish minimum test scores was provided.

Our Evaluation

The OPM description of the job analysis process (SME panels combined with a survey of SMEs) is consistent with recommended practice (as outlined in our Stage 1). MARSOC and OPM took steps to ensure that the personnel most knowledgeable about the job (i.e., job incumbents) would be heavily involved, and they had processes in place to confirm the accuracy of the resulting information with supervisors who may have additional relevant insights into the job requirements.

The remaining processes outlined in the statement of work were far less detailed. This makes it difficult to judge whether the results would provide sufficient support for MARSOC's selection processes. With respect to our Stage 3 (test validation), although OPM indicated it would take one of two acceptable approaches to validation (content validation and criterion-related validation), it did not provide any details about the approaches. We did not know what tests MARSOC or OPM would identify as relevant to validate (our Stage 2), nor did we know what data might be obtained in support of the validity process or how those data would be analyzed. So, while both content and criterion-related validation approaches are considered consistent with recommended practice for supporting a selection process, we could not determine whether or what kind of validation study would be conducted.

OPM briefly mentioned establishing minimum standards (our Stage 4) in the OPM statement of work; however, the process that would be used to establish the minimums was not described, and the work plan appeared to focus exclusively on establishing minimum performance standards for the job tasks identified in the job analysis. There was no mention of how that information would be used to tie into the establishment of minimum screening standards (e.g., test score cutoff points).

Lastly, there was no mention of whether any women would be included in the validation process or whether gender bias would be explored during the process of content or criterion-related validation, or the setting of minimum standards.

As a result of the limited information available from the OPM statement of work, there were large gaps in our understanding of the work that OPM would be doing for MARSOC. Some or all of those gaps might be eliminated with additional information that might be included in OPM's final report. The details not provided in the statement of work, such as what types of data were ultimately obtained, the statistical properties of those data, how the data were analyzed, and the conclusions that were drawn from the results, all matter for determining whether the use of specific screening criteria and specific minimums are justified.

CHAPTER EIGHT

Navy Special Operations Forces

Five Navy occupations, known as the "Warrior Challenge" occupations, require PSTs to ensure personnel can meet the physical demands of the occupations. Only two of these occupations—the special warfare operators (SEALs) and special warfare combatant-craft crewmen (SWCCs), also known as special warfare boat operators—are closed to women under DGCAR. These closed occupations are highly specialized special operations forces jobs, which together account for around 3,000 positions: about 2,000 SEALs,[1] and about 1,000 SWCCs (including both active-duty and reserve service members on three Special Boat Teams in Coronado, California; Little Creek, Virginia; and Stennis, Michigan). The remaining three Warrior Challenge occupations—explosive ordnance disposal (EOD) technician, Navy diver (ND) and aviation rescue swimmer (AIRR)— were already open to women at the time of this study and had women on the job. This chapter discusses entry standards for the SEALs and SWCCs and the Navy's ongoing activities for establishing gender-neutral selection standards for these two special operations occupations.

Occupational Assignment and Screening

The paths to entering these occupations differ according to whether interested applicants are currently serving, have previously served, or

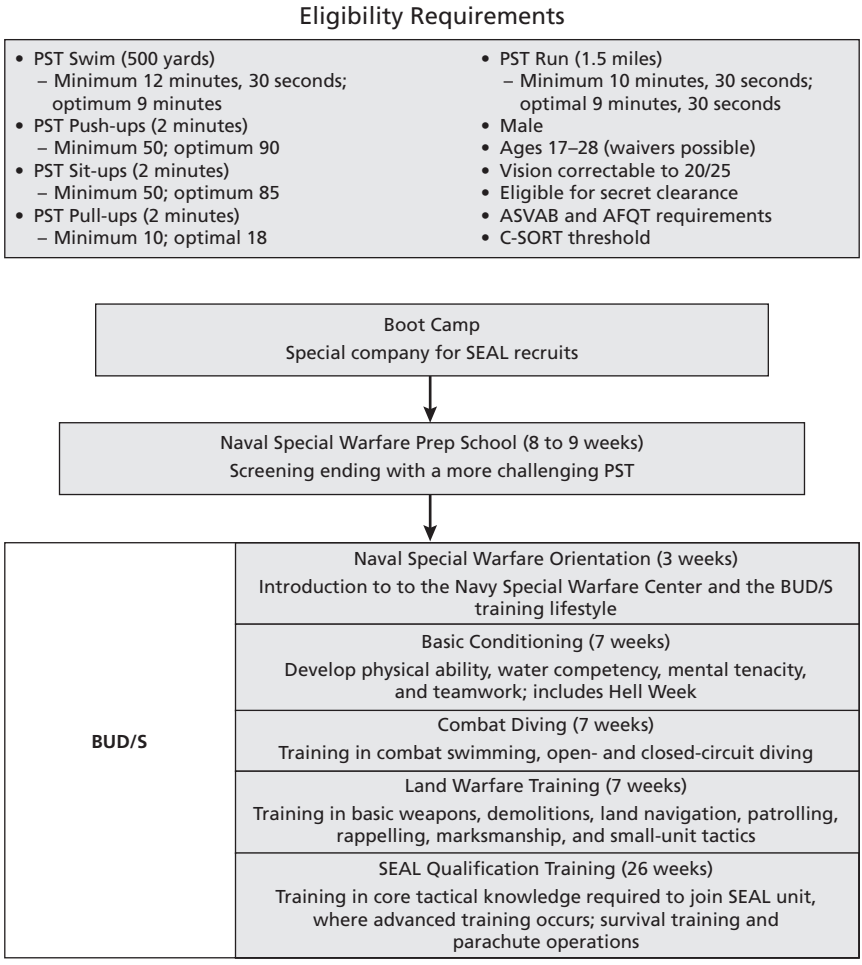
[1] Nine active-duty teams, including four on the West Coast, four on the East Coast, one SEAL Delivery Vehicle Team, and two Reserve SEAL teams.

have never served in the U.S. military, as shown in Figure 8.1. Current Navy service members can apply to transfer into these occupations by passing the standard PST for Navy personnel and notifying their command through a Special Request Chit. Many applicants, however, are individuals who have never served. Those who are new to the military and those seeking to join the Navy after previously separating from the military follow similar paths, beginning at a local MEPS. Final selection for entry into SEAL and SWCC training is made using a top-down process.

At the MEPS, recruits complete a medical prescreening report, take the ASVAB, provide documentation to demonstrate their eligibility to join the Navy (proof of age, U.S. citizenship, financial viability, education, etc.) and undergo a complete physical. Next, recruits choose their occupational specialty, with guidance from a Navy career classifier. If they choose one of the two special operations forces occupations, they must take and pass the PST and meet all other entry requirements at this point. Contingent upon meeting those eligibility requirements, recruits are considered for selection into the training pipeline for that occupation. Recruits must pass the PST again just prior to the start of boot camp to ensure they maintain high levels of physical conditioning. Although PST minimums are specified for each occupation (shown in Figure 8.1), applicants typically will need much higher scores to be competitive for selection into training for special operations forces occupations. For example, although minimum swim time to apply to be a Navy SEAL is 12½ minutes, the Navy reports that a nine-minute swim time is considered ideal. Similarly, 18 pull-ups and 90 push-ups—well above the minimums required—are considered ideal.

There are typically more people who meet the minimum standards than there are available spaces in training. As a result, each occupation can be more selective in choosing candidates for training. The PFT scores are among the factors considered in making those selection decisions, and each occupation makes final selection decisions differently. New recruits who make the cut proceed to a seven- to nine-week boot camp after signing their contracts. Officers proceed to Officer Candidate School or Officer Development Schools for five to 12 weeks.

Figure 8.1
Eligibility and Training Requirements for Navy SEALs



SOURCE: Naval Military Personnel Manual, 2013; Naval Military Personnel Manual, 2016; U.S. Navy, Recruiting Command, 2011.
NOTE: BUD/S = Basic Underwater Demolition/SEAL.
RAND *RR1340/2-8.1*

Current service members proceed directly to occupational training, as long as they again pass the PST shortly before the beginning of that training. After boot camp, the paths of recruits for each of the two occupations diverge.

For enlisted jobs, in addition to minimum scores on the PST, both the SWCCs and the SEALs require minimum scores on the Armed Forces Qualification Test (AFQT) and a set of combined ASVAB subtests. Candidates also are required to complete the Computerized Special Operations Resilience Test (C-SORT), which assesses three areas: performance strategies (goal-setting, self-talk, emotional control); psychological resilience (acceptance of life situations, ability to deal with cognitive challenges and threats); and personality traits. The three areas are combined into a score on a scale of 1 to 4 (1 is lowest). Selection criteria for officers are similar to those of enlisted.[2] SEAL and SWCC eligibility requirements are summarized in Figures 8.1 and 8.2, respectively.

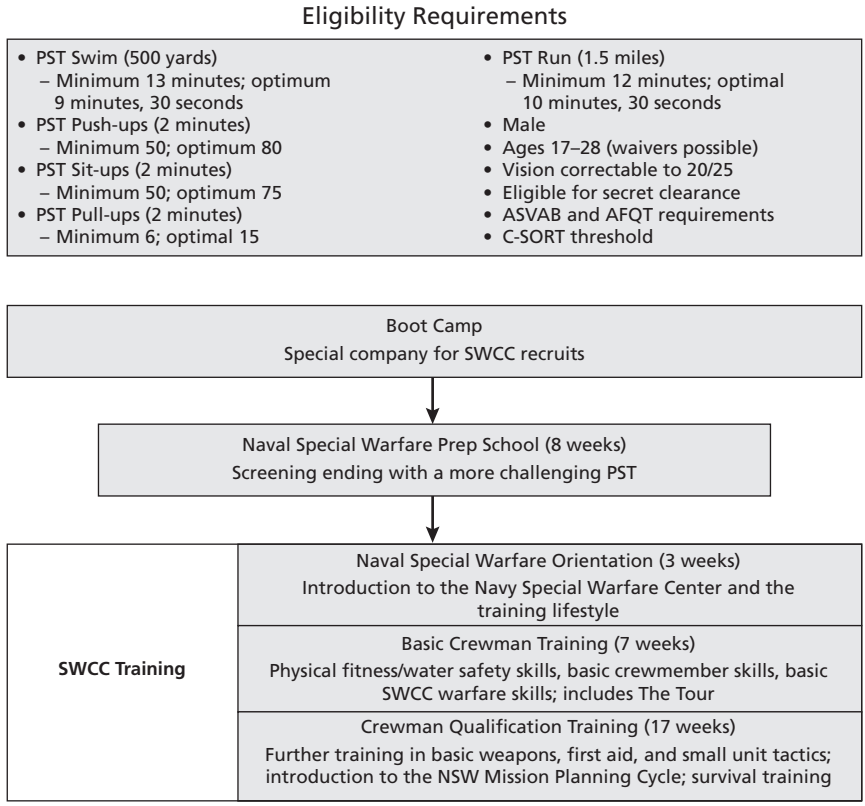
Because there are typically many more personnel interested in the jobs than there are spots in Basic Underwater Demolition/SEAL (BUD/S) training at the Naval Special Warfare (NSW) Preparatory School in Great Lakes, Illinois, only those judged to be the most competitive candidates based on their PST, ASVAB, and C-SORT scores and other screening test information are accepted. Using this information, experienced SEAL or SWCC personnel rank-order all applicants, and the top candidates are selected to fill the number of available training slots. For both of these occupations, combined PST results using an established formula are used to rank-order applicants.[3] We do not

---

[2] SEAL and SWCC officers must be commissioned from the U.S. Naval Academy, the Naval Reserve Officer Training Corps, or the Officer Candidate School. SEAL officers usually spend five years as platoon-level leaders and then move on to plan larger-scale operations in such roles as department head or training officer in charge. Most senior roles include SEAL team commanding officers. There are typically 70 to 90 officer-training slots. Successful training officers work in a special operations forces command that includes multiple services with at least 200 people. Criteria for SEAL officer selection include proven leadership, strong language/cultural expertise, strong academic performance, and highly competitive PST scores.

[3] The composite formula equals (run time in seconds + swim time in seconds) – [(# pull-ups x 6) + # push-ups + # sit-ups]. A researcher originally developed the formula, though we do not have access to that researcher's documentation in support of it. The formula is used only to determine SEALs' and SWCCs' composite PST scores; it is not applied to other physically demanding Navy occupations. According to conversations with NSWC personnel, no plans are in place to evaluate or revise the formula prior to opening SEAL and SWCC positions to women.

Figure 8.2
Eligibility and Training Requirements for Navy SWCCs



SOURCE: Naval Special Warfare Center, 2016.
RAND *RR1340/2-8.2*

know how the PST is combined with the other information to make a selection determination. We also were told that candidates with low C-SORT scores and low run and swim times are generally advised that they are not ready to pursue BUD/S training.

Once selected for training, the candidates still have to successfully complete the training pipeline. Because so many trainees fail to complete training (due to both involuntary and voluntary attrition), making it through training is a major selection hurdle for these occupations.

SEAL Training

Annually, approximately 1,000 personnel start the BUD/S training at the NSW Preparatory School, but only about 200 to 250 complete it.[4] BUD/S has six stages. Each stage includes challenges that push the candidates physically and a physical assessment that is administered to trainees, with scores recorded.

The first stage, known as BUD/S Prep, is a two-month screening training that begins and ends with a more challenging PST than the one required for entry. At the start of BUD/S Prep, trainees take the Human Performance Program (HPP) Combine test, which includes a standing long jump (assessed in inches), maximum number of pull-ups while wearing a 25-pound vest, maximum number of bench presses of the recruit's body weight, maximum number of 1.5 times their body weight in dead lift, an agility run (measured in seconds), a 300-yard shuttle run (measured in seconds), a 3-mile run, and an 800-meter swim with fins. The ending PST requires a 1,000-yard swim with fins in 20 minutes or less, at least 70 push-ups in two minutes, at least 10 pull-ups in two minutes, at least 60 curl-ups in two minutes, and a 4-mile run in shoes and pants in 31 minutes or less.

The second stage, BUD/S Orientation, is a three-week course introducing students to the NSW Center and the BUD/S lifestyle. The purpose of the course is to prepare students for the first day of BUD/S Basic Conditioning, known as the First Phase. The NSW Orientation assessment is administered during this stage. It is a physical ability assessment that consists of a core endurance test (measured with total time in seconds to hold a side plank and single-leg bridge[5] on each side) and a 1½-mile run.

The next stage (Stage 3 of the training pipeline) is where BUD/S begins. Known as Basic Conditioning, this is a seven-week course dedicated to developing physical ability, water competency, mental tenac-

---

[4] SEAL Officer Assessment and Selection takes place from May through August of each year in Coronado, California. The process includes physical screening, psychological evaluations, behavioral assessments, and team activities in a competitive environment.

[5] A side plank and a single-leg bridge are two common exercises used to test and strengthen core muscles.

ity, and teamwork. Weekly tests include a 4-mile timed run, a timed obstacle course, and a timed 2-mile swim. Hell Week, consisting of physical training for more than 20 hours a day with only four hours of sleep spread over 5½ days, takes place during the fourth week. Many candidates ask to drop out of SEAL training during this phase. This voluntary attrition accounts for a large part of the selection at BUD/S.

The second phase of BUD/S (Stage 4 of the overall training pipeline), known as Combat Diving, is a seven-week course providing instruction on combat swimming and open- and closed-circuit diving, both of which are skills unique to SEALs, and the third phase (Stage 5), Land Warfare Training, focuses on skills related to basic weapons, demolitions, land navigation, patrolling, rappelling, marksmanship, and small-unit tactics.

The final training stage and the last stage of BUD/S, SEAL Qualification Training (SQT), is a 26-week training course in core tactical knowledge. It provides survival, evasion, resistance, and escape preparation, as well as advanced training in weapons, small-unit tactics, land navigation, demolitions, cold-weather training, medical skills, maritime operations, and static-line and free-fall parachute operations. Upon completion of the SQT, trainees receive their SEAL Trident, designating them as SEALs. They are then assigned to a SEAL team and usually receive an additional year of training before their first deployment.

SWCC Training

Each year, approximately 240 personnel start the 8½-month Basic Crewman training for SWCCs at the NSW Preparatory School in Great Lakes, Illinois, but only about 120 complete it. Basic Crewman training has four stages.

The first stage, NSW Preparatory School, like the BUD/S Prep course, is a two-month screening training program that begins and ends with a more challenging PST than the PST required for entry. The ending PST during NSW Preparatory requires a 1,000-yard swim with fins in 22 minutes or less, at least 50 push-ups in two minutes, at least seven pull-ups in two minutes, at least 60 sit-ups in two minutes,

and a 3-mile run in shoes and pants in 24 minutes or less. If candidates fail this test, they are reclassified to other Navy jobs.

The second stage, NSW Orientation, is an introduction to Basic Crewman physical training, with a focus on running, swimming, pushups, sit-ups, pull-ups, and obstacle course performance. Basic Crewman Training (Stage 3) is a seven-week course with a focus on physical conditioning, water competency, teamwork, mental tenacity, basic navigation, and small boat seamanship. The physical intensity and mental intensity of the activities increase each week and culminate with The Tour—a three-day application of skills and physical abilities with limited sleep that fills the role that Hell Week does in SEAL training.

Finally, Crewman Qualification Training (Stage 4) is a 17-week course in which candidates progress to intermediate levels of seamanship and navigation, weapons, communications systems, marksmanship, engineering, waterborne patrolling, mobility, and combat casualty care. The training also includes an introduction to the NSW Mission Planning Cycle of preparing, planning, briefing, and executing an NSW mission, as well as the Survival, Evasion, Resistance and Escape course and continued physical training.

Navy's Process for Validating Existing Standards for Special Operations Forces

The SWCC and SEAL occupations are under SOCOM oversight, but the Naval Special Warfare Command (NSWC) had responsibility for validating standards, and the work itself was designed and carried out by researchers at the NHRC. The remaining sections in this chapter describe the evidence gathered by NHRC to support the validity of the SEAL and SWCC process and the work done to establish gender-neutral standards. This work focused on the training standards that serve to select which personnel can continue and enter the occupations. NSWC relied on existing research to support the PST requirements that determine eligibility to apply and are used to rank-order applicants

for selection into training. We first review this existing research before discussing the NHRC effort.

Studies of Existing Candidate Selection Criteria

NHRC's data collection efforts did not include a reexamination of the PST requirements for screening to decide who can enter training in the closed occupations. NHRC instead directed us to more than 30 existing studies about SEAL and SWCC screening and training criteria conducted since the 1970s as support for the PST. The studies cover a wide variety of topics, not all directly pertinent to the establishment of gender-neutral physical standards. For example, several of the studies focused on finding temperament and personality measures for use in predicting who will be successful in BUD/S (and Hell Week in particular). Others examined use of AFQT and ASVAB composite scores for predicting attrition. A few studies do, however, provide a direct examination of the relationship between PST scores and training attrition.

The first relevant study was based on a large sample of fairly recent data that was analyzed by NSWC personnel in 2014.[6] NSWC explored PST and training attrition data from more than 7,000 BUD/S students. The data show that only 27 percent of all trainees completed the training and that there is a relationship between PST composite scores and training success. These results suggest the existing highly competitive screening process (which includes consideration of PST scores) is not doing a particularly good job of identifying which candidates are most likely to succeed in training.

The researchers recommended an increase to the minimum PST standards for application to be a SEAL based on the observed relationship between PST scores and training success and the need to improve training success. However, how much the minimums should be increased is not specified, and alternative ways to improve physical screening are not explored. For example, the data do not show strong relationships for all of the PST test components. The average number

[6] Although the study was dated 2014 (after DGCAR was lifted), it is not clear whether the study was conducted in direct support of the work to respond to the NDAA language or for other reasons.

of pull-ups was the same for officers who failed and who passed (17), suggesting no relationship between this requirement and success. For enlisted personnel, the number was 12 for those who failed and 13 for those who passed, again suggesting essentially no relationship. This raises questions about whether the composite score formula and the elements within it should be reexamined before making adjustments to the score minimums.

In addition, there are many other unanswered questions about these data and the findings. For example, the data show that the average PST scores for successful officers were quite different from those for successful enlisted personnel (e.g., mean swim times for successful officers and successful enlisted personnel were 9:00 minutes and 9:55 minutes, respectively), but the reasons for this and the implications are not discussed. Perhaps the standards for officer performance during training are set higher than for performance of enlisted personnel. Or perhaps enlisted personnel are able to make greater gains in performance during training than officers, but the bar for performance during training is the same for both. Either circumstance raises issues for how to set entry standards. Therefore, both are of direct concern in marshaling support for the use of the PST.

The study recommends implementation of a tiered minimum composite PST system, requiring a 1,200 score to enter into the SEAL challenge contract or delayed entry program, followed by a required 1,100 to enter recruit training and enter NSW prep. Recruits would then need to score 1,000 to enter BUD/S with an eventual goal to be determined through competition based on quotas. However, no evidence in support of such a tiered system is provided. Given the current method for selecting from the applicants, it is unclear what effect raising the score minimums would have on training success (or subsequent performance on the job). At the time of this study, applicants who scored at or close to these minimums have essentially no chance of being accepted.

NSW Basic Training Command conducted a second study related to the PST. Although the exact date of the study is unknown, it used data from 2012 and 2013 SEAL (n = 291) and SWCC (n = 86) training classes. The study explored the relationship between the three physi-

Table 8.1
Test Scores of SEAL and SWCC Graduates and Nongraduates

Test Name	Test Elements	SEAL		SWCC	
		Nongraduates	Graduates	Nongraduates	Graduates
HPP Combine	Standing long jump (inches)	89.3	91.5	87.7	89.3
	Max. pull-ups with 25-lb vest	10.9	11.8	8.4	9.1
	Max. reps bench press @ body weight	10.5	10.5	8.2	7.7
	Dead lift reps @ 1½ x body weight	4.8	4.7	4.3	4.6
	5-10-5-yard agility run (seconds)	4.92	4.84	4.9	4.9
	300-yard shuttle (avg. of two attempts with 2-min recovery, in seconds)	62.4	61.3	63.3	62.7
	3-mile run (hours)	0.86	0.81	0.9	0.87
	800-meter swim with fins (hours)	0.59	0.58	0.60	0.58
NSW Prep	1,000-meter swim with fins (hours)	0.73	0.72	0.75	0.72
	Max push-ups in 2 minutes	82.7	86.7	72.7	74.3
	Max sit-ups in 2 minutes	76.6	80.4	72.3	77.8
	Max pull-ups	15.6	15.9	13.5	13.8
	4-mile run/3 miles for SWCC (hours)	1.17	1.12	0.89	0.88

Table 8.1—Continued

Test Name	Test Elements	SEAL		SWCC	
		Nongraduates	Graduates	Nongraduates	Graduates
NSWO	Core endurance test: side plank and single-leg bridge on right and left sides (score is total time in seconds)	548	590	515	555
	1½ -mile run data for BUD/S students (hours)	0.41	0.40	0.43	0.41

SOURCE: U.S. Naval Special Warfare Command, 2014.

cal performance assessments conducted during the first two stages of training and overall success in training. Table 8.1 shows results of the three assessments. For SEAL and SWCC recruits, those who completed training scored better on several elements of the earlier physical assessments. For example, the results for the NSW Preparatory exit test and the NSW orientation assessment also show that those who completed training performed better on every task than did the drops. This data support the conclusion that the earlier physical tests administered in training generally relate to training success, which is heavily influenced by completion of Hell Week and The Tour.

Although none of the aforementioned studies included female participants, the Navy Recruiting Command in 2013 reported data on average male and "top performing" female PST scores among Navy Challenge recruits who sought to join one of the physically demanding Navy occupations (all five occupations for men and the three open occupations for women). The data showed the top woman scored lower than the average man on all the physical tests. A more complete analysis of gender differences, however, would show the distributions of scores on each test for men versus women, relative to the scores of successful applicants.

Our Evaluation of the Navy Studies of Candidate Selection Criteria

How the PST scores affect the rank orderings by the reviewers and how the rank orderings correlate with subsequent performance will need to be investigated to determine whether the screening process is valid. If women are permitted to apply in the future, this investigation must evaluate whether these relationships are gender-neutral, i.e., whether the role of PST scores and other factors in the rank ordering of applicants is equal for men and women and whether the rankings predict training success in the same way for both genders.

There are many gaps in the past research that still need to be filled to provide support for the use of PST scores for selection into training or PST scores for continuation in training. Because PST scores are used in part or whole to rank-order candidates for selection into training, there needs to be strong evidence showing not only that higher scores are associated with performance in training and subsequently on the job, but also that the process for ranking applicants uses the scores appropriately. There also needs to be evidence that the scores predict equally well for both men and women, something past studies do not explore.

Evidence exploring which of the subcomponents of the PST are predictive also is needed. Revisions to the formula for creating a composite score likely would be needed based on the results of such research. For example, none of the prior studies to which we have access provides strong support for the predictive or content validity of the pull-ups portion of the PST. Such evidence would be needed to justify its continued use.

Study to Validate Gender-Neutral SEAL and SWCC Training Standards

As we indicated earlier, NSWC turned to NHRC to conduct the research for establishing gender-neutral standards in response to lifting of DGCAR. In documentation provided to us and in our discussions, NHRC described the following steps in an investigation of the extent to which SEAL and SWCC selection requirements during training are related to occupational performance. The approach included updating the job analysis listing operational and mission essential tasks to reflect

current and anticipated future SEAL/SWCC mission demands, surveying job incumbents to identify individual attributes likely to predict performance, validating key training events, and linking testing early in training to subsequent training outcomes.

Identify Physical Demands

As one of its primary data collection efforts, NHRC updated an older job analysis for the SEALs conducted in 1995 (Prusaczyk et al., 1995) and conducted a new job analysis for the SWCCs. The approach used methods similar to the 1995 study, which included soliciting SME input in defining in-theater scenarios and conducting a survey of job-incumbents asking for a variety of information about the in-theater scenarios.

To develop the scenarios, NHRC held focus groups with 112 SEAL and 64 SWCC noncommissioned officers (participants were E-6s and higher with an average of six deployments). NHRC asked the SMEs to describe a variety of in-theater scenarios that characterized typical on-the-job SEAL and SWCC activities. The scenarios were written as sets of realistic tasks during typical missions and included such details as equipment used and weights of objects used. With the help of additional SMEs, the list was narrowed to eliminate redundancies. Finally, NHRC conducted several surveys of the same types of job incumbent SMEs[7] to finalize the scenarios.

The first survey asked SMEs to rate the following for each of the mission scenarios:

- How physically difficult is it to perform the mission relative to all other missions?
- How important is it for SEALs or SWCCs to be able to perform the mission relative to all other missions?
- How frequently is the mission (or one very similar) performed compared with all other missions?

[7] Participants for each survey may have included the same SMEs and new SMEs.

Identify Physical Attributes Likely to Predict Performance on the Job

Other questions in the surveys described earlier were focused on defining which attributes are needed on the job. One set of questions asked participants to check off each personality and physical attribute that they believed was relevant to success in the mission scenario. Examples of the physical attributes included aerobic fitness, upper- and lower-body strength and endurance, core stability, coordination, and strength and power. Each attribute was defined for the survey participants.

Another set of questions asked participants to rank-order a list of attributes from 1 to 20, according to importance for successfully completing the mission sets. The attributes were a set that had been previously identified as relevant for success as an operator, including maturity, professionalism, tactical professionalism, integrity, humility, creativity, conduct, leadership, teamwork, confidence, discipline, situational awareness, aggressiveness, and strength.

The NHRC researchers suggested in an early draft report that they planned to use this information to establish the content validity of the training standards, but no explanation of how the data would be used to accomplish this was provided. Again, we were not able to see the results of these survey items or how they would be analyzed.

Validate Key Training Events

As we discussed earlier, Hell Week and The Tour screen out many SEAL and SWCC trainees. To validate these events, NHRC sought to evaluate the link between them and mission performance on the job. To do this, NHRC relied on expert judgment by including questions in its survey for SEAL job incumbents to provide their judgments as SMEs about Hell Week or The Tour.

One set of survey items asked participants to check off the physical training activities during Hell Week that they felt were critical or useful for preparing them for success in each mission scenario. In a preliminary briefing of the results, all Hell Week evolutions (except the life story) were rated as useful for operational performance by more than 95 percent of participants, and about half were rated as essential by more than 90 percent of participants. However, we did not have the

exact wording of the item or a description of the methods for analyzing the data from the survey.

An additional set of survey items also contained questions to support the link between Hell Week and performance on the job, including

- Have you ever experienced a situation during an operation that was as challenging as Hell Week?
- How frequently during operations do you experience situations as challenging as Hell Week?
- Did you gain greater confidence in your own ability to overcome challenges as a result of completing Hell Week? By how much?
- Did your confidence in the abilities of others increase as a result of them completing Hell Week successfully? By how much?

Preliminary results in response to these questions were characterized by NHRC as demonstrating the validity of the Hell Week training content. The results appear to show strong beliefs among participants that Hell Week builds confidence in themselves and others, and that nearly all participants had experienced a situation during an operation that was as challenging as Hell Week.

Although the NHRC researchers concluded that Hell Week is valid based on the survey results, some of the past studies that we reviewed (dating back to the 1970s) raised questions about whether Hell Week content is justified. For example, studies noted that instructors can be idiosyncratic in how they treat each student (Bretton and Doherty, 1979; Doherty, Trent, and Bretton, 1981; Hoffman, 2002). In addition, many studies have noted the extremely low pass rates, even among the most physically fit and prepared candidates (for examples, see Breton and Doherty, 1979; McDonald, Norton, and Hodgdon, 1990; Aleton et al., 2002; Mills and Held, undated; Peterson, Pihl, and Pihl, 2006; Hoffman, 2002; Mills and Robson, undated; San Diego State University, 2002; and Doherty, Trent, and Bretton, 1981). Some have asked whether many of those who were driven to quit actually would have been successful on the job (see, for example, Doherty, Trent, and Bretton, 1981; Breton and Doherty, 1979). Senior leaders in the Navy also suggested there was a need to better connect Hell Week

requirements to the job requirements, well before DGCAR was lifted (The Thomas Group, 2006). In response to these issues, researchers have recommended standardizing the training across instructors and better connecting the training activities to on-the-job requirements (see, for examples, Bretton and Doherty, 1979; Doherty, Trent, and Bretton, 1981; and Hoffman, 2002).

Link Testing Earlier in Training to Later Training Outcomes for SEALs

In the preliminary findings provided to us, NHRC presented correlations exploring whether physical performance at different stages of SEAL and SWCC training and testing are related. Each of the BUD/S First Phase physical activities (underwater swim, drown-proofing, obstacle course, timed swim, Hell Week, lifesaving, knot tying, pool competency, and treading) was correlated with each of the physical testing activities (log physical training, land portage, rucksack march, down man drills, sand bag physical training, timed run, surf passage, rock portage, paddling, surf immersion, boogie man swim, knot tying, and swimming in surf). NHRC provided similar relationships related to comparable SWCC training. The researchers concluded on the basis of the data that SEAL and SWCC training outcomes reflect physical testing scores.

Unfortunately, the information we received did not describe the source of these data or the process used to arrive at these scores. Assuming that the scoring process was sound, this analysis could indicate that performance in earlier training events is related to performance in later selection events—therefore, it could be used to screen out trainees long before they begin Hell Week or The Tour. However, without additional information about how the scores were derived, we cannot evaluate the relevance of the correlational data.

In the same preliminary findings, the NHRC researchers report a relationship between PST run and swim times at entry to the NSW prep course and the PST run and swim times at the end of NSW prep.

Our Evaluation of the Navy's Approach to Validating Existing Training Standards

In updating the job analysis (Stage 1 in the standard approach to setting standards), NHRC used data collection methods consistent with the

approach in the previous SEAL job analysis and similar to approaches typical in job analysis. We do not know whether respondents were asked whether any important missions were missing, and we were not able to see the results of the responses to these items or how they were analyzed. We also did not have access to documentation of the many technical elements that lend support to the job analysis. This includes agreement among SMEs, confirmation by SMEs that important information was not missing, and information about how reflective some of the details included in the descriptions (such as weights of equipment) are of actual mission demands.

Identification of individual attributes likely to predict job performance aligns with Stage 2 of the recommended standard setting process. Job incumbents served as the SMEs who provided judgments regarding the linkages between the attributes needed to be successful and the mission descriptions. It is not clear to what extent the SMEs agreed about the linkages or to what extent outside observers would agree. Job incumbents are not necessarily experts at understanding personality traits or physiology and, as a result, their judgment about which attributes or how much of the attributes are needed could be called into question. Additional evidence that multiple outside experts (such as researchers in physiology and personnel psychology) would arrive at the same independent assessment would have strengthened the conclusions. Collecting physiological measurements and conducting observations of people successfully performing the activities would establish even greater support for the conclusions.

NHRC's assessment of the link between the content in training and the content on the job is entirely based on perceptions by job incumbents in responses to a few items on a survey. Job incumbents believe Hell Week helped them gain confidence in themselves and others, that it is essential to preparing a candidate for operational performance, and that they have faced equally difficult situations on the job. Although this is evidence that can be marshaled to provide some support for that link, that support could face very legitimate criticism. For example, some of the questions the researchers asked require a logical leap that may not be justified with the current data. Although job incumbents reported having faced an equally challenging situation to

Hell Week on a mission, similarity in judged difficulty alone is insufficient to show that the Hell Week content is reflective of the content on the job—the tasks must also be the same or highly similar. The survey did not collect information on task similarity. Empirical links between confidence gained from Hell Week and successful performance on the job were also not provided. It is possible that confidence gained from Hell Week could be achieved in other ways, which might not be accompanied by attrition of large numbers of personnel. Although few would argue that confidence is irrelevant to successful performance in these occupations, such confidence might be attained through other activities than the ones performed during Hell Week. The research did not evaluate alternative approaches. In addition, the NHRC analysis did not explore how consistent the training difficulty is from person to person and class to class, a concern raised by earlier studies. Therefore, this is still an area worthy of further investigation.

Finally, we cannot provide a complete assessment of the NHRC analysis to validate the physical testing early in training (related to Stage 3 of the recommended process) without knowing more about how the correlations were computed and other specifics about the data. Nevertheless, based on our preliminary understanding, it appears that there could be potential gaps in the logic for why these relationships should matter. Finding a correlation between performance early in training and later in training is relevant only if passing and failing training at each point is clearly tied to success or failure on the job. The fact that PST scores, which are highly abstracted from the job (i.e., not directly emulating actual work activities), correlate with themselves at a later time demonstrates that they are reliable at rank-ordering trainees on the activities tested. However, it does not prove that this rank-ordering predicts how well the trainees would do on the job.

Our Evaluation

Selection into the SEAL and SWCC occupations begins when applicants are initially selected for training and continues through the entire training period. Applicants must have minimum PST scores, but ini-

tial selection is based on a rank-ordering process that combines the PST scores (which are well above the minimums for application) with other information. The recent studies of PST scores are not adequate for validating either the PST as a physical screen for selection or its use in rank-ordering applicants.

Acceptance into BUD/S or SWCC training is the first of several selection points before candidates enter these occupations. The multiple stages of training involve increasingly challenging physical assessments and training activities, culminating in the ultimate screening during Hell Week and The Tour. Almost all of those who continue beyond this point will graduate and enter the occupations, but a high fraction of those initially selected drop out of training up to and during these two events.

Because Hell Week and The Tour serve as the final selection point and the level at which many trainees drop out, the NHRC research for NSWC focused on whether the activities during these events reflect mission requirements on the job and whether scores on the physical assessments earlier in training predict performance during the two events and other training activities. When the researchers conclude that SEAL and SWCC standards are valid, we assume that they are referring to these physical assessments and training standards during BUD/S and The Tour rather than the standards for initial selection for training. Based on the information available to us, we offer several observations:

- The research linking training content to job requirements is based on a content analysis that relies on the opinions of job incumbents. No data were collected to empirically validate whether performance in Hell Week and The Tour predicts performance on the job. Such data would go a long way toward furthering support for the continued use of the two training events as they now stand.
- We could not evaluate the analysis done to determine whether the sequence of physical assessment tests predicts training performance and completion without more documentation of the methods and data used.

- None of the data collected by NHRC included female participants because there were no women currently on the job or in training. Therefore, it is unclear whether the relationships reported between earlier training testing performance and later points in the training would be the same for women as for men. This is an area that should be explored further using data on women in training or female research subjects.
- The NHRC work has not explored the validity of the PST and its use in rank-ordering applicants for entry into SEAL/SWCC training. The past studies of the PST find relationships between some, but not all, elements of the PST and training attrition.
- NHRC has not collected data appropriate for establishing minimum levels of performance that should be considered passing during Hell Week and The Tour, during the other training blocks, or on the initial PST prior to entry into training (i.e., Stage 4). It is not clear to us how the existing minimum standards have been set or whether and how they will be revised to meet the NDAA requirements.

Importantly, our review was based on an early draft of the write-up of the methodology for the NHRC work. That write-up lacked the detail and clarity necessary to fully understand and critically evaluate the approach NHRC researchers took in the work. Much of it focused on discussing the importance of presenting a persuasive argument for validity instead of on the actual work conducted to support that argument. It did elaborate on some of the data collection efforts, but again, the details and rationale for the processes were incomplete. Some of the potential gaps we have identified in this chapter may disappear with a more detailed and complete description of the research and how it was ultimately used to set valid, gender-neutral standards.

CHAPTER NINE

Air Force Battlefield Airmen

Only seven occupations—as well as the associated units and training courses—were closed to women because of the combat exclusion policy. Personnel in these occupations (both officers and enlisted) are collectively known as *battlefield airmen*. The seven are

- special tactics officer (STO)—13CX
- combat control team (CCT)—1C2X enlisted
- special operations weather team (SOWT)—1W0X2 enlisted
- special operations weather team (SOWT)—15WXC officer
- pararescue (PJ)—1T2X enlisted
- combat rescue officer (CRO)—13DX
- tactical air control party (TACP)—1C4X enlisted.

Although these occupations are all physically demanding, the jobs are quite different. CCT personnel oversee air traffic control in austere environments. SOWT personnel serve as meteorologists who provide information to inform mission planning, route forecasts, and special reconnaissance. TACP airmen coordinate air support of ground combat and clearing of airspace. PJ personnel are trained as emergency medical technicians to operate in humanitarian and combat contexts on conventional and unconventional rescues by air, land, and sea. CROs also coordinate and directly engage in rescuing people and resources. STOs lead and coordinate the work of the PJs, CCTs, SOWTs and TACPs. Personnel in these occupations often serve as integral members of Army Ranger and Navy Seal teams and must be physically prepared to perform as members of those teams.

As of April 2013, these Air Force specialties (AFSs) together accounted for 4,686 positions closed to women (Donley, 2013). The rest of the more than 500,000 positions that exist in the Air Force are open to women.

Occupational Assignment and Screening in the Air Force

Entering enlisted personnel are assigned to a career area at time of enlistment by counselors at the MEPS based on a variety of factors including ASVAB scores, physical aptitude test scores, and Air Force needs. In a majority of cases, specific occupations are not guaranteed to personnel at that time. However, some harder-to-fill occupations are assigned prior to enlistment and guaranteed to the candidate in the enlistment contract. In those cases, the Air Force guarantees that personnel will not be reclassified into a new occupation as long as they continue to meet requirements for the job. Those who do not meet the requirements (by failing or withdrawing from training, for example) are reclassified into a different occupation. Which occupations carry such guarantees can vary from year to year. For those without a guaranteed occupation, assignments to an occupation happen during basic training using data from the MEPS to determine eligibility.

The majority of Air Force Specialty Codes (AFSCs) for officers have no physical requirements for entry into training. However, many enlisted AFSCs do have physical requirements. Those that do have such requirements rely on scores from two physical aptitude tests: the physical ability and stamina test (PAST) and the SAT. The SAT has been in place for decades, and the PAST was added more recently to address attrition from battlefield airmen training pipelines. Both were established by exploring relationships between test scores and performance—the PAST by predicting who washes out from training and the SAT by predicting laboratory simulations of physically demanding job activities.

Enlisted recruits are eligible to pursue battlefield airman occupations from time of recruitment, as long as they meet the eligibility conditions. Therefore, unlike other special operations forces, selection for

battlefield airman training is based on minimum screening test scores and does not involve a top-down process. In addition to being male, the selection conditions include (see Figure 9.1) meeting the PAST and SAT minimums; U.S. citizenship; eligibility for at least a secret security clearance; maximum age of 28 (though age exceptions are made for those with prior military service); passing the standard military physical and the Class III Flight Physical; and fulfilling specific vision, height, and weight requirements. Officers face similar eligibility conditions for the battlefield airman occupations.

The SAT is used for screening enlisted personnel only. All enlisted applicants must demonstrate the ability to lift 40 pounds on an incremental lift machine prior to enlistment. This minimum requirement is met by nearly everyone who applies to enlist in the Air Force. However, many enlisted occupations have higher incremental lift requirements that range from 50 to 100 pounds for a select few occupations—and many applicants do not meet those higher requirements. Those who score lower than 100 are eligible for fewer occupations.

Enlistees who do not meet the specified required minimum incremental lift score for entry into a given occupation are barred from that job. Minimum scores higher than 40 pounds are a requirement for entry into several occupations currently open to women. Among enlisted battlefield airman occupations, SOWT requires a minimum incremental lift of 50 pounds; CCT, TACP and PJ require 70 pounds.

Career counselors to all enlisted applicants at the MEPS administer the SAT. Historically, a sizeable proportion of women and a much-larger proportion of men achieve at least a 70 on the SAT (the average score for women is around 71, whereas the average score for men is higher than 100). Given this, many women and most men achieve scores that meet or exceed the battlefield airman SAT minimums. For more on the SAT, see Sims et al. (2014).

The PAST has been used for several years to prescreen personnel for entry into enlisted battlefield airmen occupations in the Air Force. It is also used to screen for entry into two other physically demanding

occupations already open to women.[1] The PAST includes a 25-meter underwater swim (assessed as either pass or fail), a timed 500-meter surface swim, a timed 1½-mile run, a count of chin-ups completed in one minute, and push-ups and sit-ups completed in two minutes. Recruits can earn a range of points for each of the events, with a total possible score of 330; passing requires a 270 or higher. However, as shown in Table 9.1, minimum scores on each individual event must also be met, and those minimums differ by occupation.

Table 9.1
Physical Ability Stamina Test Minimums for Enlisted Jobs

	Eligibility for Women	Underwater Swim (2 x 25 meters)	Max. Time for 500-m Swim (minutes)	Max. Time for 1.5-Mile Run (minutes)	Minimum Pull-ups	Minimum Sit-ups	Minimum Push-ups
PJ	Closed	Pass/fail	10:07	9:47	10	54	52
CCT	Closed	Pass/fail	11:42	10:10	8	48	48
SOWT	Closed	Pass/fail	14:00	10:10	8	48	48
TACP	Closed	N/A	N/A	10:47	6	48	40
EOD[a]	Open	N/A	N/A	11:00	3	50	35
Survive Escape Resist Evade trainer[a]	Open	N/A	10:00 (200-m swim)	11:00	8	48	48

SOURCE: "Air Force Special Tactics Fitness Training," 2011; HQ AETC/A3T homepage, undated.

[a] EOD and Survive Escape Resist Evade trainer are already open to women.

[1] For more specifics on the PAST, see the Air Force Officer Classification Directory (AFOCD) (U.S. Air Force, 2013b) and the Air Force Enlisted Classification Directory (AFECD) (U.S. Air Force, 2013a).

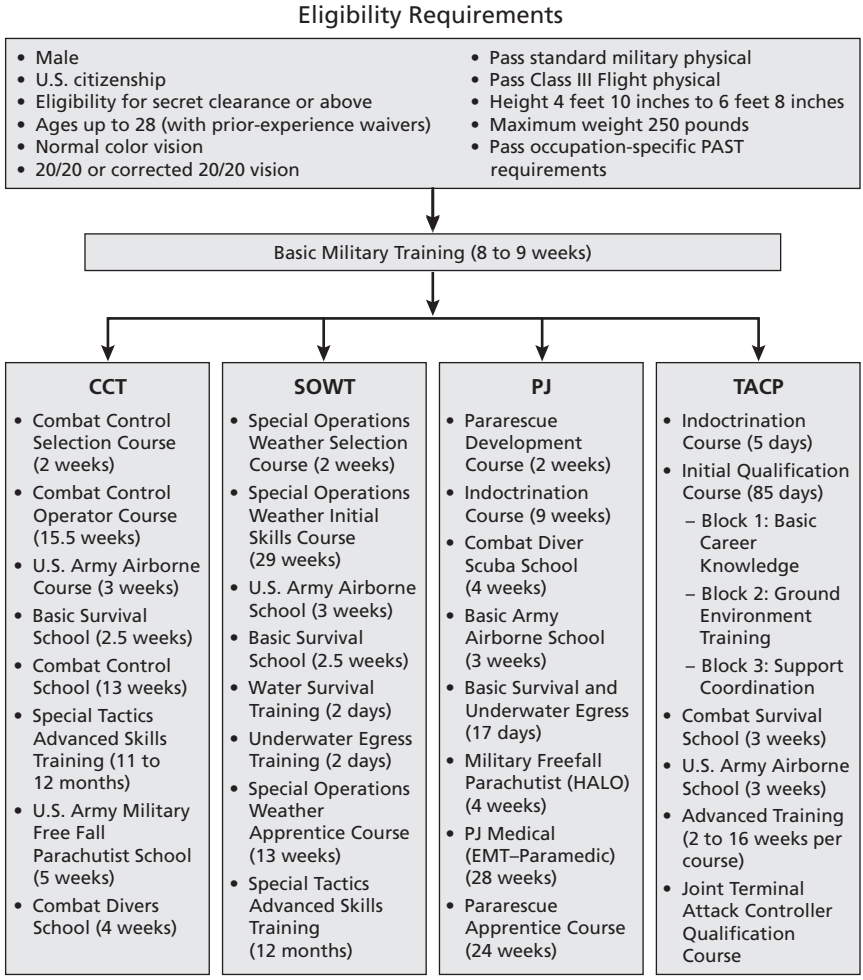
For enlisted applicants, the PAST is administered at least three times to determine eligibility. It is initially administered by recruiters to those individuals who are interested in one of the battlefield airman occupations. It is not administered to all Air Force applicants. Those who pass the PAST are given the test on three more occasions: before shipping to basic training, in the first week of basic training, and again during the transition between basic and technical training.

For STOs and CROs (the battlefield airman officer jobs), a modified version of the PAST (with slight variations in the content and ordering of the test) is administered as part of the initial application for entry into training. For example, the STO test requires one 25-meter underwater swim instead of two, and it requires a 1,500-meter swim rather than a 500-meter swim. For those events identical to the enlisted PAST (such as pull-ups), the minimum scores differ for officers.

All enlisted recruits attend basic training and then proceed into occupation-specific technical training. Those who do not meet minimum expectations for completing each block of training are either washed back (i.e., allowed to begin the training block again with the next incoming class), or they are washed out of the training entirely and reclassified into another occupation. For battlefield airman occupations, there are a variety of academic and physical challenges that serve to wash out and wash back trainees. As a result, training can serve to further narrow the training pool, and there have historically been high washout rates from the training in these occupations (see, for example, Manacapilli et al., 2012). All the battlefield airmen occupations include some sort of assessment and screening block specifically designed to narrow the pool of trainees. For PJs, for example, this winnowing occurs during the PJ indoctrination course. Figure 9.1 shows the key training blocks for each of the seven occupations closed to women.

The Air Force also has annual fitness standards for some of these occupations. For example, personnel in SOW, STO, and CCT occupations must demonstrate that they meet minimum fitness standards on a series of physical fitness tests (including number of sit-ups, chin-ups, and run and swim times) upon arrival at their first assignment and again periodically throughout their career. Lastly, a new operator

Figure 9.1
Eligibility and Training Requirements for Enlisted Battlefield Airmen



SOURCE: U.S. Air Force, "U.S. Air Force Special Ops," webpage, undated.

RAND *RR1340/2-9.1*

test for battlefield airman occupations recently has been established to ensure the standards required during training also are being maintained in the operational force.

Air Force Process for Establishing Standards for Battlefield Airmen

The Air Force's Director of Force Management Policy, Deputy Chief of Staff for Manpower, Personnel and Services (HQ AF/A1P), is the office with primary responsibility for establishing gender-neutral standards for closed occupations. A1P delegated the planning and implementation of that work to the USAF Fitness Testing and Standards Unit under Air Education Training Command (AETC/A1), which provides exercise physiology science consultation to the Air Force Deputy Chief of Staff for Manpower and Personnel and to AETC/A1 on the forcewide fitness assessment program and policy. Although the unit is conducting many elements of the research and data collection effort, the Air Force also commissioned an FY 2014 RAND study for some elements of the research. The work to develop new battlefield airman standards began in 2011.

The Air Force explored new physical aptitude tests to screen people for entry into battlefield airman specialty training pipelines and ultimately replace the PAST. The result would be a new set of occupation-specific physical screening criteria (referred to as Tier 2 fitness standards).[2] The following sections describe the steps in the process the Air Force used to identify and validate the new battlefield airman screening criteria. The information was gleaned largely from our interviews with the RAND researchers and the USAF Fitness Testing and Standards Unit, and from the documentation provided to us by the USAF Fitness Testing and Standards Unit (including unpublished briefings and written study plans).

Job Analysis

The first step in the research effort involved a detailed job analysis to define the critical physically demanding tasks in each job. The job analysis process started with a series of focus groups in which multiple groups of SMEs (consisting of five to eight senior noncommissioned

[2] Tier 1 standards are intended to ensure the general health of the force and therefore are applicable regardless of occupation.

officers and officers in the occupations) were convened to review and refine preexisting task lists provided by the Air Force's Occupational Analysis Flight within AETC.[3] SMEs were sampled to ensure representation within each job by mission, unit, and environment and were required to have had at least one operational deployment within the past five years.

The task list, narrowed to only those involving physical activities, was presented to participants. For each task, participants were asked to provide an example to describe the activity; rate its frequency, importance, intensity (using a 1-to-5 Borg scale), and duration; and detail the physical actions used during the activity (pull, press/push, bend, squat, lift, crawl, climb, etc.) They were also asked to provide relevant information, such as combat loads, distances traversed, whether it was a team or individual activity, mechanical advantages, and environmental conditions. Based on this information, AETC created a final physical task list. That list then served as the foundation for a survey of all airmen in the battlefield airman specialties in which participants rated the tasks on the same set of dimensions. The final job-specific lists of physically demanding tasks were compiled from the survey results. Only those tasks that were rated as both physically demanding and critical to the job were included on the list. This final list was then presented to a panel of senior leaders and more junior personnel to determine (1) what proportion of personnel should retain that capability, and (2) whether any tasks from other operational environments were missing from the list.

[3] Official task lists are produced by the Occupational Analysis Flight and updated every three years (or more frequently when changes to the job warrant it) for all enlisted occupations. Task lists are developed by first soliciting SME input to confirm the relevance of the existing tasks from the prior year's lists and identify gaps. Then, all personnel are surveyed and asked to rate each task that was confirmed or added by the SMEs. Task lists for officer occupations are developed only on special request, but Occupational Analysis Flight follows similar procedures for developing them when requested. Task lists resulting from Occupational Analysis Flight's survey and SME inputs are compiled into an official report, which is made available to the career field managers and training developers. These official task list reports were used as the starting point for the SME focus groups described earlier.

Criterion-Related Validation Study to Replace the PAST

Data collection for the criterion-related validation effort was slated to begin in April 2015. In that effort, AETC/A1 identified and administered a range of physical tests for use as potential predictors, and it designed and administered a range of physically demanding job performance simulations for use as measures of job performance. The researchers estimated that, when complete, the tests and simulations would be administered to a sample of at least 200 personnel—50 job incumbents (all men) and 150 tech-training students from other careers (including about 80 women). To prevent fatigue effects, data collection was designed to take place over a period of two weeks to allow for scheduled rest days and break times between tests and simulations. During the two-week testing window, participants were introduced to the simulations and tests and provided training and practice opportunities to ensure they were familiar with the activities and knew what was expected of them before testing. Week 1 was designed to focus on the screening tests; week 2 focused on the simulations.

The Air Force completed a pilot of all of the simulations and the physical test battery in March 2015. The goal for the pilot study was to smooth out unanticipated data collection problems (such as equipment issues and lack of variance in participant scores), refine key test administration features (such as testing times, distances of rucksacks, height of walls, repetitions, and appropriate weight loads), verify that key activities were still judged as appropriate and realistic by job incumbents, and reduce the list of tests and simulations to a more manageable number. Pilot participants included battlefield airman trainees and job incumbent SMEs. Participation from others was also solicited, as needed, to further test equipment and protocols. Multiple rounds of testing and refinement took place during the pilot.

The researchers used a systematic process for deciding which candidate screening tests to include in the pilot. They started by examining roughly 600 physical tests explored in past research (including published research in military and nonmilitary contexts). They composed a matrix, where each test was rated on certain criteria—feasibility (i.e., ease of implementation), cost, validity, risk of injury, liability, and others. Roughly 60 of the tests (including those showing the great-

est promise from the matrix) were chosen for inclusion in the pilot study. Among the tests under consideration were the PAST elements and related variants (such as weighted pull-ups and weighted push-ups). Using the results of the pilot test, the number of tests was further narrowed. Only a subset of the original 60 was selected for inclusion in the full data collection effort. The tests retained for the full data collection effort had not been finalized by the conclusion of our data collection period in April 2015, but they were expected to include a variety of different types of screening tools, with slight variants on each type.

The simulation activities chosen included a series of land, tower (simulating climbing activities), and water-based tasks designed to emulate critical physically demanding performance tasks identified during the job analysis. The primary measures of performance in the simulations were time to complete each activity and/or total distance completed within an allotted time frame.

Some of the simulations were isolated and short in duration. For example, one was intended to simulate a boat carry over obstacles across a beach area before putting the boat back in the water. Participants picked up a bag with a handle (meant to simulate the boat) and walked through pea gravel the distance of a typical beachhead. Another simulated a rock climb by having participants climb up a wall and then pull up their rucksacks. Others, however, were combined into a realistic sequence to elicit the same physical task conditions (including fatigue) as would be faced on the job.

The small unit tactics simulation, for example, was the most complex and time consuming of the simulations. It started with a 5-kilometer ruck march after which participants completed the following activities in sequence: a low crawl 24 inches high (Army standards); a buddy drag; an evasion maneuver with cones and obstacles (various wall heights); a maneuver over an 8-foot wall with a 2-foot bench assist; a fireman's carry up and down a flight of stairs; a sled drag followed by a jog (repeated twice to simulate a team task where one runs while the other drags); a litter carry up a ramp; and a litter lift.

To determine appropriate distances, weights, speeds, and other requirements for the scenarios, researchers collected data (such as heart-rate and weight of the rucks) from experienced operators while

they were executing full mission profiles (i.e., realistic mission scenarios designed to emulate real battlefield conditions and terrain) in the United States. Executing these full mission profiles is a regular part of maintaining battlefield airman operator currency. They take place in a variety of locations (e.g., Florida, Alaska, Hawaii, Colorado) to simulate different climates and terrain. AFSOC assigns operators a realistic mission set to complete, and the full mission profile is built from that mission. Battlefield airman SMEs with relevant field experience develop the details of the mission. Researchers identified key design features for the simulations from the data collected at all of the full mission profile training locations.

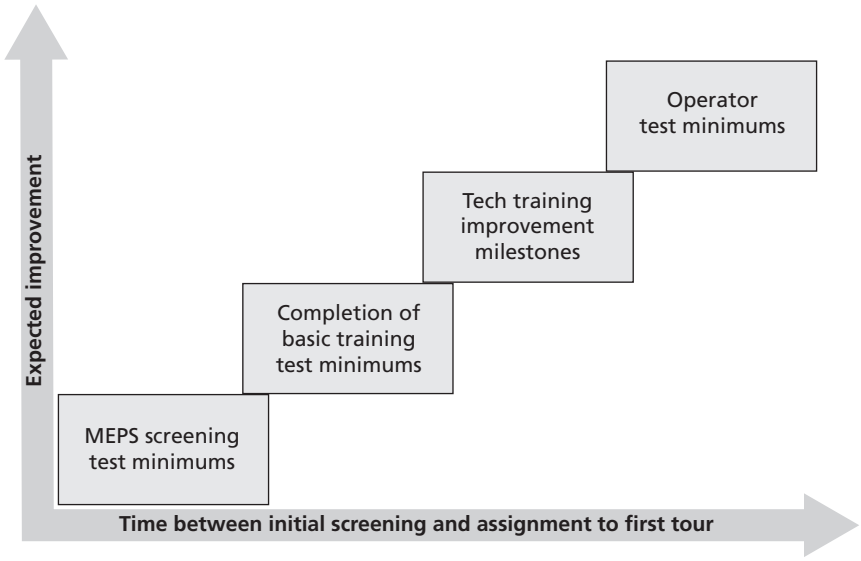
The plan called for all participants to complete all tests and participate in all simulations; however, not all of the simulations were relevant for all of the battlefield airman occupations. Only those simulations relevant to an occupation were used to establish its requirements.

How the Criterion Validation Results Were Used

The criterion-related validation study data was first used to establish the recommended annual testing standards (such as how many pull-ups) for the battlefield airman operators (the people currently performing the job). In other words, the study aimed to identify the tests and minimum scores that best determine who is physically ready to perform on the job. The resulting minimum scores on those tests (specific to each occupation) would then be used to certify each current battlefield airman's capability annually. Once the tests and score minimums were determined, they were to be proposed to leadership for concurrence.

After the operator tests and minimum scores are established, AETC/A1 planned to work in reverse over the training pipeline to establish the training entry requirements. As shown in Figure 9.2, entry requirements used to qualify personnel for training would be designed to account for improvement and development over the training period. How much improvement to expect would be estimated using data from published research on physical aptitude training and archival data on male trainees in past battlefield airman training pipelines. Working backward from the desired performance end-state (i.e., minimum scores on the operator test), the analysts would use the

Figure 9.2
Incrementally Higher Minimums Account for Improvement Gained from Training



RAND *RR1340/2-9.2*

improvement information to estimate the starting requirement needed at time of entry at the MEPS location and at completion of basic training (the two key selection decision points in determining eligibility for entry into the training pipeline described earlier).

Because training pipelines differ in length, the amount of improvement that can be expected would differ by occupation. Thus, even if operator requirements are the same across occupations, the training entry requirements could differ. The longest battlefield airman pipeline is for PJ, which takes about two years. Because that pipeline is so long, significant gains in physical ability are likely prior to entering the occupation. For other battlefield airman career fields where the pipeline is shorter, the potential for gains necessarily will be lower. As a result, the researchers acknowledged that having different training entry standards for two distinct career fields, even if the final physical demands on both occupations are the same, is a very real possibility.

How the Operator Test Minimums Would Be Established

The researchers first set minimum standards for performance on the simulation activities. To establish those minimums, a subset of the 50 criterion-related validation study participants who are experienced operators would participate as SMEs in a standard setting panel. The panel was to take place during the criterion-related validation study. Those SMEs selected for the panel (i.e., those identified by the career field as having the appropriate experience and expertise) first completed a simulation themselves as part of the normal study data collection. After completing the simulation, they were told how they performed on it (e.g., how fast they completed it) as part of their role on the panel. They were then immediately asked to identify the minimum level of performance that would be expected by someone considered minimally competent in the job.

The results of the SME standard setting panels were shown to leaders and compared with actual operator performance on the simulations as a final check on the accuracy of the performance minimums established by the panel participants. Once minimums on the simulations were identified, minimum test scores would be determined using the statistical correlation between the scores and the simulation activities on a subset of the participant sample. The remainder of the sample would be used as a hold-out sample to cross-validate the minimums. That is, the statistically derived test minimums would be applied to the hold-out sample, and the amount of error in predicting who was successful in the simulations would be explored. Test minimums could be revised depending on the results.

Once the operator tests were selected and minimums set using this process, the researchers planned to complete one more final check of the minimum test scores by having experienced operators complete the tests and then execute full mission profiles as part of the existing operator practice events regularly conducted in the United States. This last step was intended to allow Air Combat Command and AFSOC to verify that the established standards work as intended in the operational environment. Those who can meet the screening test standards should also be able to perform the relevant tasks in the operational environment; those who fail to meet the test standards should not per-

form satisfactorily in the operational environment. For example, if a person meets the test standard that predicts the successful completion of a given time and distance on a task simulation involving a rope ladder climb, he or she should also successfully complete a rope ladder climb in a realistic mission setting where conditions (such as wind, rain, darkness, fatigue, etc.) might differ from those in the static testing environment. If this can be confirmed in that final step, it provides further support for the use of the test.

The test score minimums that result from this process were to be used to define the operator test minimums. As explained, the test scores required at earlier points in the career (e.g., during training or upon entry to the service) were to be adjusted to account for expected improvement during training.

Our Evaluation

The Air Force's job analysis methodology (focus groups with SMEs to develop and refine a task list, a follow-on survey using the task list, and then final confirmation of the findings with SMEs to ensure that important tasks are not missing) is consistent with the practices we outlined in Stage 1 of the recommended practices. The job analysis results should serve as a good foundation for the later steps in the validation process, but again, details matter. We did not receive a write-up of the findings from the job analysis data collection process, so we could not determine whether the data analyses and conclusions drawn by the researchers were sound.

The Air Force also explored a wide variety of screening tests for inclusion in the study, which is consistent with what we recommended in our Stage 2. They used a well-reasoned and systematic process for initially narrowing the list of potential screening tests (described in earlier sections) and then included a sampling of tests in the validation study with the intention of using the study findings to further narrow the test list. Their process focused on using empirical findings to drive the final test content, which again is consistent with recommended practice.

With respect to our Stage 3, an important element to consider in a study such as the Air Force's (i.e., a simulation-based criterion validation study) is how well the simulations actually reflect the requirements of the job. The Air Force designed its simulations based on the job analysis results and on realistic details collected from full mission profiles used by current battlefield airman operators during their ongoing training. Assuming that the full mission profiles are accurate reflections of the conditions under which the activities might be performed, their use in combination with the job analysis findings lends strong support for the content validity of the simulation activities. However, again without final documentation, we could not confirm how well the simulation activities emulated the real circumstances under which personnel are required to perform. The extent to which the simulations capture all relevant physical dimensions of performance (i.e., are not construct deficient), show consistency in individuals' performance (i.e., test-retest reliability), and elicit the appropriate level of difficulty are just a few of the important details in evaluating the final results.

Other important factors to consider with respect to Stage 3 include whether the types of data collected on the simulation activities and the predictor tests are appropriate, and whether the statistical analyses run on that resulting data are appropriate. Close examination of the resulting data on both the tests and the simulation activities would reveal whether the test scores demonstrate the appropriate statistical properties (including sufficient variance). The resulting regression findings and the conclusions the researchers draw with respect to those findings also are important.

With respect to our Stage 4, the Air Force articulated a plan for how it planned to establish minimum scores on the tests that is also consistent with recommended practice. Officials first planned to use job-incumbent SMEs to identify the level of simulation performance that would be expected of a minimally competent person in their occupation. The proposed steps to check the SME judgment against the judgment of others knowledgeable about the career field are a strength

of the methodology;[4] however, if the double-checking by others suggested changes were needed, there would need to be a strong rationale and sound justification for those adjustments. The next step was to use the relationships established in the criterion validation effort to crosswalk the resulting minimum simulation performance levels to corresponding minimum levels on the predictor tests. Lastly, the Air Force planned to use estimated training gains over time to establish the minimum test scores required to initially qualify for training.

In sum, it appears that many elements of the Air Force's criterion validation effort are consistent with recommended practice. The researchers took steps to collect solid data on which to base their decisions at important points in the validation process, and they planned to have data supporting many of the important links that are critical in a well-designed criterion validation study. This is apparent from the existing documentation and the interview information provided to us prior to our study ending. However, the formal write-up of the methods, analyses, and findings were still forthcoming when our study ended; therefore, many fine-grained details of the data analysis decisions were unknown to us. In addition, although there are strengths to the approach that can lend credibility and support to any resulting test score minimums, we did identify some potential gaps in the work. In particular, examination of bias of the testing by gender was not addressed in the plans described to us. Although the plans did not include examination of bias by gender, the collected data could still be used to do so.

---

[4] Note that subjective judgment can in some cases be highly questionable; however, in other cases, it can be a useful contribution to validation evidence. When multiple SMEs with appropriate expertise are included, and evidence is collected to show both the reliability and the accuracy of SME judgments, the use of subjective judgments can provide a sound methodological approach.

CHAPTER TEN

Overarching Observations, Findings, and Recommendations

Comparing the Services' Efforts

The services took different approaches to amassing evidence to develop and support their screening standards. Differences in the approaches do not mean that one effort is better than the others, as there are always multiple sound options for how to approach the work. Nevertheless, those differences will have bearings on what conclusions can be drawn from each of the respective efforts. Table 10.1 summarizes the approach taken in each case, and some of the more notable differences are discussed next.

Quality of the Empirical Support for the Screening Standards

Although differences alone are not necessarily a sign that one effort is better than another, there were some qualitative differences in the soundness of the conclusions that could be drawn as a result of the services' respective efforts. In particular, the researchers designing the Army's combat arms effort and the Air Force's battlefield airman effort generally pursued methodologies that were well suited to addressing the first four stages of our recommended six-stage process. Their investigations relating to each step were carefully designed to provide sound data and solid links between each research element and each step. As a result, based on our review of the work they completed by the end of our study, these two efforts generally provided greater empirical support for the validity and utility of their screening processes relative to the other efforts described in this report. In sum, while the rest of the

Table 10.1
Summary of Key Features of the Service Approaches

Service	Selection Process Being Validated	Stage 1 Identify Physical Demands (Job Analysis)	Stage 2 Identify Potential Screening Tests	Stage 3 Validate and Select Tests
Army combat arms	Screening before training	Review of existing job analysis materials through SME interviews, focus groups, and incumbent survey to rate frequency, importance, and time spent	Twelve candidate predictor tests, chosen to measure types of physical abilities identified by SMEs as needed for physically demanding tasks	Concurrent criterion-related validation to determine how well candidate tests predicted performance on simulated job tasks
Army Special Operations Forces	Training	New in-depth job analysis by OPM using occupational information, site visits, job incumbent survey	Existing training activities	Content validity, details to be determined
Marine Corps combat arms (Phase 1 study)	Screening before training	Job tasks identified from existing training and readiness manuals, which rely on occupation-specific task lists regularly updated based on SME review and a job incumbent survey	Elements of existing PFT and CFT	Concurrent criterion-related validation to determine how well candidate tests predict performance on basic physical tests roughly similar to physically demanding job tasks

Table 10.1—Continued

Service	Selection Process Being Validated	Stage 1 Identify Physical Demands (Job Analysis)	Stage 2 Identify Potential Screening Tests	Stage 3 Validate and Select Tests
Marine Corps combat arms (Phase-2 study)	Not clear how results will be used to set standards	Unit mission events developed by SMEs representing multiple Marine Corps organizations including operational combat organizations	Data collected included an unknown number of potential screening tests	Concurrent criterion-related validation to determine how gender mix of a unit and individual physical characteristics affected unit performance and, to a lesser extent, individual performance during unit events
Marine Corps Special Operations Forces	Training	New in-depth job analysis by OPM using occupational information, site visits, job incumbent survey	Existing training activities	To be determined
Navy Special Operations Forces	Training	New job analysis with SME input and job incumbent survey; also developed mission scenarios using focus groups of experienced job incumbents and incumbent survey to determine difficulty, importance, frequency of mission scenarios	Existing training activities (Hell Week in particular)	Content validity through job incumbent judgments of attributes relevant to success in mission scenarios and relevance of Hell Week to actual operations, identified through survey of job incumbents

Table 10.1—Continued

Service	Selection Process Being Validated	Stage 1 Identify Physical Demands (Job Analysis)	Stage 2 Identify Potential Screening Tests	Stage 3 Validate and Select Tests
Air Force battlefield airmen	Screening before training	Job analysis with review of existing task lists by SME focus groups and survey of job incumbents; final review by panel of senior and junior incumbents	Identified new tests based on test criteria determined in the research literature, pilot study of 60 candidate tests	Concurrent criterion-related validation to determine how well candidate tests predicted performance on simulated job tasks

services' efforts have marshaled some form of support for their screening process, much of that support is—to varying degrees—less definitive than that of the efforts for the Army combat arms and Air Force battlefield airmen.

Operationalizing "Physical Screening"

Each service conceives of its physical screening in a slightly different way, and, as a result, the work to validate the physical screening processes had a somewhat different focus. The Army and Marine Corps work for ground combat occupations will be used to establish gender-neutral standards for selection into these occupations at entry. The Air Force's efforts for its special operations occupations were focused on the same objective. In contrast, the work by the Army, Navy, and Marine Corps for their special operations occupations focused heavily on validating the training content and did not validate the use of PST results in the top-down process used for initial applicant selection. However, in each case, the information obtained through the research can be useful for informing the validity of the other screening elements. For example, the Army designed a simulation-based criterion-related validation study for the ground combat occupations in which individual-level task simulations were designed, measured, and analyzed with attention to detail and data were collected linking them to

more-realistic occupational task requirements. Although the focus was on validating screening criteria, the process of identifying critical operational task requirements could help support the validity of the training content. Similarly, the Marine Corps undertook two studies for its ground combat occupations, one designed to modify its fitness test for use in selecting recruits for entry into ground combat occupations and the other to relate measures of individual physical capacity to performance in simulated unit activities. Like the Army's work, the information from the simulated unit activities could be useful for informing training validity, although that was not the stated goal of the research. Lastly, the Air Force's work validating the initial screening criteria for use with its special operations jobs included an especially thorough job analysis, which helps clearly define the requisite training content.

Similarly, although the work for the remaining special operations occupations in the Navy, Marine Corps and Army focused on validating the training content, the work could serve as a starting foundation for work validating selection criteria. For example, NSWC's work focused on validating the training content exclusively, because that is where much of the intensive screening takes place through voluntary and involuntary attrition. That work, however, could help suggest new screening criteria that should be considered and explored in a validation study. The Army and Marine Corps special operations forces work by OPM, in contrast, appeared to be designed to inform whether the contents of the training and the screening tools are relevant. The planned work focused heavily on conducting a thorough job analysis, which has the potential to be used not only to validate training content (the clearly stated objective of the research), but also to lay the foundation for a validation study of screening criteria (another potential element of the OPM work, although a methodology for such an effort had not yet been clearly articulated at the end of our data collection period).

Comparing Highly Similar Jobs Across Services

Differences in the services' efforts are likely to receive especially close scrutiny for jobs that appear to be highly similar across services. Infantry jobs, for example, will be a natural comparison to make between the Army and Marine Corps efforts. Because the two services took very

distinct approaches to establishing standards for infantry, it would not be surprising if they end up with somewhat different screening criteria. The methodological differences need not imply disparities in the validity of the screening criteria; that is, both could be equally valid. However, if differences in screening criteria lead to greater adverse impact for women in one selection process than the other, or if one leads to a much greater number of personnel being excluded (i.e., much higher standards), then it will be useful to identify the legitimate reasons for the differences to exist. For example, if the screening process in one service happens much earlier in the career than in the other service, lower physical screening minimums may appropriately reflect added training time during which personnel improve their physical conditioning before the first job assignment. Or, although the two jobs may share the same name, it is possible that Marine Corps requirements differ from Army requirements, thus justifying differences in the screening criteria.

Establishing Occupation-Specific Versus Combat Arms–Specific Standards

The Marine Corps is the only service that designed a study to establish a single standard for all of its ground combat occupations. This appears to reflect a legitimate difference in both the culture of the organization and the way in which the members of these occupations may be utilized. TECOM motivated its approach by the observation that all members of its combat forces (regardless of specialty) must be capable of meeting the physical demands of any combat arms occupation. This anticipates that Marines in combat units may be called upon to perform in any of the combat arms occupations and therefore must be prepared and capable to meet those duties at all times. The other services, however, have not taken such an approach. They have established standards for each occupation that are specific and applicable to that job only.

Remaining Stages of the Standards Setting Process

Service Processes for Establishing Minimum Acceptable Scores

When we completed this research in April 2015, the services had not yet begun to establish minimum selection standards, and only the Air Force had identified a method for doing so. This stage (Stage 4) is key to determining whether the standards are set appropriately. If they are set too high, people who are capable of performing on the job will be unfairly excluded. This could affect the mission, as people with other characteristics needed for performance on the job (e.g., intelligence, persistence, mechanical or language skills) might be being excluded unnecessarily. The people chosen may be stronger, but they may be less capable in other ways. It could also lead to an inability to fully man an occupation if not enough people qualify. On the other hand, if the standards are set too low, the mission could also suffer because some of those selected would not be capable of performing required duties. As a result, setting the bar for the minimums is a critical step in the process of establishing standards. Detailed documentation on this stage in the process will be critical to supporting the use of the screening minimums chosen.

Service Processes for Addressing the Implementation Stage

Implementation (Stage 5) will also be key. Many things could occur during implementation that could invalidate the screening for predicting who will be successful. Administration of the tests in a manner that is inconsistent, incorrect, or different from the way the tests were administered during the research; whether and how selectees prepare for the tests; availability of retesting; and when the tests are administered are just a few examples of ways that implementation could undermine the validity of the tests in practice. There should be a plan in place to ensure that these issues have been carefully considered and any potential for inconsistencies guarded against. The services should continue to monitor their implementation procedures to ensure they are being followed and no unanticipated changes have occurred that could result in reduced validity.

Service Plans for Continued Research After the Standards Are Implemented

Stage 6 recommends that research on the standards continue long after the standards have been implemented. This is to address the many potential naturally occurring gaps in research to validate standards, and it is a vital step to amassing evidence to support the continued use of any screening standards. In the case of setting standards for the closed occupations, there are several gaps in the research that can be best addressed in Stage 6.

For example, although the services' efforts all have been huge undertakings, many critical research questions cannot be answered until the selection processes are in place and administered operationally. It will be important to follow up after implementing the standards to see whether the standards have good predictive validity in practice. Essentially, how good are they at distinguishing the good performers from the inadequate ones? However, note that the ability to explore such relationships later could be severely limited by what is known as *range restriction*. If nearly everyone selected performs well on the job or in training, or if the people chosen have only the highest scores on the selection criteria, then there will not be enough variance in test scores to observe a relationship. This is a common problem when examining selection processes already in place. These issues would need to be accounted for when examining how tests are functioning after implementation. There are statistical approaches that can be used to help account for range restriction issues in some cases; however, in cases where range restriction cannot be corrected statistically, other methodological approaches for confirming validity and exploring bias will need to be explored.

Reexamination of the findings on a regular basis is also an important process for ensuring that the validity of the screening does not change over time. Jobs change, technology changes, test administration practices can change, and even the population itself can change (e.g., women and men could be better physically prepared before even applying for the job). Therefore, the Office of the Secretary of Defense (OSD) should explore what the services have planned to accomplish this regular reexamination of the validity and fairness of the screening

criteria. The services should establish policy that puts systems in place to address these issues in a systematic way and on an ongoing basis.

Finally, for all of the services, there were some unavoidable limitations in what could be completed prior to opening the positions. One issue in particular that was unavoidable in all of the services' work was that no women are actually in the closed jobs yet. As a result, there was no pool of incumbent women for the researchers to draw upon as research participants. The only SMEs with deployed experience performing the job are men. Women participating in simulation activities (such as in the Marine Corps, Army, and Air Force efforts) did not have operational experience comparable to the male counterparts. This omission of job-experienced women poses an unavoidable dilemma. Because the NDAA mandated that evidence supporting the validity of the standards be in place prior to opening the jobs to women, the services were unable to validate the standards on a real female applicant and job incumbent pool. Such a pool will take years to develop and normalize. Applicants and applicant qualifications likely will change as people adjust to the opening of the positions and it becomes a more accepted career path for women. As women become interested, they will undoubtedly begin to prepare in earnest to meet the physical demands. As a result, we recommend continuing to collect data on the validity of the screening criteria and alternative measures on samples of both men and women applicants and incumbents in the years following the opening of the positions. Institutionalizing ongoing data collection to support replication of Stages 1 through 4 periodically[1] to

[1] How frequently each stage should be repeated depends on how often jobs and available screening tests change and how quickly women enter the occupation. Parts of the job analysis process (Stage 1) should be replicated at least every several years to confirm that the job content has not changed. However, if there is reason to believe a job has changed in the intervening time, that should trigger a new job analysis much sooner. Replicating the validation process (Stage 3) should occur frequently when new tests are instituted. The tests should be revalidated as soon as selection and performance data on training classes and performance information can be amassed, and again when sufficient data on female applicants becomes available. This will allow further refinement of the tests and the amassing of greater evidence to support the continued use of the tests. After a variety of evidence has been amassed, it would still be important to further validate the tests every several years or every decade or so to verify that the key factors (such as length of the intervening training, administration

include examination of validity within each gender group will be essential to ensure that the tests and selection criteria are valid in the future.

Lastly, as noted in our description of best practice methods, no single research effort can address all issues, and no research study is without limitations. The limitations we were able to identify given the stage of completion of the services' work and documentation are idiosyncratic to the different research designs chosen. As always, research often raises additional questions while at the same time answering others. As a result, Stage 6 (continued research) is an important next step after the standards are in place and the jobs are opened to determine whether the physical standards are effective or the gaps we identified (or other shortcomings in the design and implementation of the standards) mean that adjustments will be needed.

In sum, because ongoing research is a critical step in setting the best standards, the public and OSD should expect to see the services continue to investigate the validity of their standards long after implementation. They should also be aware that it is legitimate for the services to continue to make adjustments to the standards informed by that ongoing research and as new information comes to light over time. In other words, the answer to date should not be taken as the definitive solution to selection in these occupations. Continued changes and adjustments to the standards should be not only expected, but also encouraged as more information is amassed.

Formal Documentation of All Aspects of the Work

If not documented properly, research on physical standards holds little value in lending support to practices over time. Details, such as the overall statistical and methodological approaches, summary statistics, data analyses, sampling approach, and participant characteristics,

---

procedures, test difficulty, and individual preparation) have not led to unforeseen changes in validity or test bias. If there is any reason to believe that the test validity has been compromised, that should trigger conducting a validation study sooner. Stage 4 should be replicated at least as frequently as the test validation process, if not more frequently, to ensure that the standards are not set too low or too high. However, continued refinement of the minimum scores by conducting ongoing standard setting studies at least initially for a few years after the tests are implemented will be critical to ensuring the standards are set appropriately.

are all necessary for experts to be able to judge the soundness of the research findings. These details also are critical for responding to any challenges to the selection practices. For that reason, we recommended in Volume 1 of this report that the services create and retain detailed write-ups of all research conducted to support and evaluate occupational physical standards. In December 2015, DoD made public the documentation for much of the research described in this report. Documentation for the remaining stages of the process—setting minimum standards, implementing the standards, and confirming the standards work as intended—should also be made public when it is available.

Final Thoughts

The call to develop valid standards was taken seriously by the services. All of the services dedicated a large amount of time and resources to their work in response to the lifting of DGCAR. As a result, the service efforts were large undertakings. Some services have set aside dedicated testing locations, simulation equipment, and scientific physiological measurement equipment. The numbers of voluntary participants joining in the work have also been impressive. Calls for participants (both men and women) have gone out to service personnel, and many have stepped up to address that call. In the Army, for example, participants had to leave their home stations and put their regular work duties on hold for weeks while they participated in the research. All told, the work that the services put forth reflects a valiant effort to accomplish exactly what was being requested: the establishment of gender-neutral valid physical standards. In addition, although the types of expertise and experience levels of service personnel involved in the efforts differed across the services' efforts, all have sought to involve personnel with a background and expertise in physiological research or personnel selection. Some services had such experts in house, whereas others sought out the assistance of experts outside of their organizations.

That said, some of the services' efforts were more comprehensive, had fewer limitations to the findings, and produced stronger evidence to support the validity of their tests. The fact that the services put forth

significant effort and resources should not overshadow the fact that some of the research efforts did not undertake all the steps needed to develop gender-neutral standards that will accurately and equitably predict performance in the newly opened ground combat occupations.

APPENDIX A

Terminology Used in Setting Physical Standards

There is often confusion in policy circles about the terminology on establishing gender-neutral standards. Many of those involved do not think to explicitly define the terms they use and typically assume the definitions are understood and shared by everyone. But sometimes the same terms are used in substantively different ways both within and across organizations. Our meetings have involved a range of service personnel, including some with substantial background relevant to personnel selection and others with limited background at best, so it is not surprising that we have observed differences in the use of key terms.

To help OEPM in its discussions with the services and clarify our use of various terms, we have prepared this appendix on terminology. Our report summarizing the recommended stages in establishing physical standards (Hardison, Hosek, and Bird, forthcoming) covers these terms in more detail, so this appendix is a summary of key points.

The Personnel Research Community Has Established Definitions

The personnel selection research community has made great strides over several decades in defining and refining the terms involved in establishing gender-neutral standards. There are many sources targeted toward academic or practitioner audiences that summarize current consensus on those definitions. The most authoritative sources, which

are published by the professional associations affiliated with the personnel selection research communities, are

- *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology, 2003)
- *Standards for Educational and Psychological Testing* (Joint Committee on Standards for Educational and Psychological Testing, 2014).

The Equal Employment Opportunity Commission and the U.S. Department of Labor have adopted these sources, so it would make sense for DoD to adopt them. Hence, the definitions we provide in this appendix are intended to be consistent with those endorsed in these two cited documents.

These authoritative sources are intended for a practitioner audience that already has some technical understanding of personnel selection but are less accessible to other audiences. They also do not address all the specific terminology issues currently facing the DoD community. Therefore, this appendix is intended to supplement the information in those sources by offering a nontechnical discussion of key terms and including only the information most relevant to establishing standards for physically demanding jobs. For broader discussion of the terms associated with selection and testing practices, we direct readers to these two cited sources.

This appendix is intended to promote a common language among the researchers involved in setting occupational physical standards in the military, and it is also intended to help ensure that DoD policymakers, Congress, and the public understand what the services mean when they describe their work to develop gender-neutral, occupationally relevant standards. Encouraging military leadership to adopt and endorse the shared definitions would be particularly helpful in establishing a consistent message on these issues.

Terms and Concepts Needing Greater Clarity

Screening, Selection, and Standards

The terms *selection* and *screening* are often used interchangeably in personnel selection. They can refer most broadly to activities occurring at any point potentially involving decisions to exclude people from entering or continuing in a job. According to this broad definition, they can include, but are not limited to, selection for specific occupations and assignment to specific jobs, wash out or wash back because of an inability to meet training standards, failure to pass a professional competency or certification test required to continue in the current job, or mastery of a new competency to continue or move up in the job.

To avoid confusion, here we differentiate these terms. Consistent with the sources cited in the previous section, in this report we use *screening* to refer to any activity that tests or measures individuals' capabilities to perform physical tasks required in an occupation. We use *selection* to refer to decisions to allow or deny entry to an occupation or, later in the career, continuation in an occupation. Thus, during screening information about individuals' occupation-specific physical capabilities is collected to support selection decisions, which are made based on *standards* set to ensure those serving in the occupation can perform at the level required to carry out the mission.

With respect to opening combat jobs to women, many have raised concerns that the initial entry standards will unfairly exclude some from an occupation or allow unqualified personnel into the occupation. Others have expressed the same concern about the hurdles that occur throughout training. Both the entry and training hurdles are of particular concern, as they are the first screening points that the first female recruits entering these occupations will face. Given limited resources and the time urgency for establishing occupational entry and training standards, it would be reasonable for the services to focus their efforts on standards for occupational entry and training standards.

Although occupational entry and training standards are arguably the most immediate concern, similar concerns could be applied to other selection points across a person's career in the service (e.g., annual physical proficiency testing). Those other selection points should also

be examined carefully to ensure that the standards are directly related to occupational requirements, and not set so high that they unfairly exclude people or set so low as to allow personnel to continue in the occupation who are not capable of satisfactory performance on the job. If these later stages cannot be addressed now, the services should include in their implementation plans how they will address the other occupational hurdles in the future.

Tests, Scores, and Measures

The terms *tests*, *evaluations*, *assessments*, *tools*, and *measures* can be used interchangeably to refer to anything that measures some aspect of a person's performance, motivation, or their underlying KSAs. Any criterion that is used to exclude or disqualify someone from a job is essentially operating as a test or a measure of that person's capabilities. Those who are excluded have, in essence, been judged to have insufficient KSAs, motivation, or performance to qualify for or continue in the job. Although many screening tools and measures are undoubtedly utilized prior to, during, and after training, they may not be recognized as such. Because of that, some could mistakenly conclude that the requirement to validate occupational-entry standards before opening closed occupations applies only to activities clearly and officially labeled as *selection tests*. As a result, the services could fail to recognize other existing types of assessments that also will need to be validated.

Scores are numerical representations of performance on a test. There are two types of numerical test scores: *criterion-referenced* and *norm-referenced*. *Criterion-referenced scores* are anchored to a specific and concrete level of performance. Getting a score of 80 for lifting 80 pounds is an example of a criterion-referenced score. *Norm-referenced scores* are defined by a comparison to performance of others. A score of 80 for lifting as much weight as the top 80 percent of test-takers is an example of a norm-referenced score. When used for selection, criterion-referenced scores can often be more straightforward to defend than norm-referenced scores. Using the earlier examples, if personnel have to lift objects weighing 80 pounds on the job, requiring a score of 80 that corresponds to lifting 80 pounds is more defensible than a score of

80 that shows they are able to lift more than 80 percent of the others who took the test.

Performance is also sometimes scored using subjective categories. Examples of such categories could include excellent, good, satisfactory, and poor; or exceeds expectations, meets expectations, does not meet expectations. When these categories are left to the rater to interpret, they are not criterion-referenced. However, if subjective labels are applied to criterion-referenced scores, then the scores and the corresponding labels can be considered criterion-referenced as well. That is, if an 80-pound lift is required on the job, lifting 80 pounds on the test could be labeled as "meets expectations," lifting 100 could be labeled as "exceeds expectations," and lifting less than 80 pounds could be labeled as "does not meet expectations." Again, criterion-referenced scores are the most defensible types for making selection decisions.[1]

Used interchangeably with the terms *hurdle*, *cut score* and *requirement*, the term *standard* in personnel selection refers to a criterion that an applicant must meet to enter or remain in an occupation. A minimum score on a physical test used to determine who is qualified for a job is one example. Standards are often defined in terms of passing/failing an established cut score or required activity. For example, trainees might be required to demonstrate a passing score on a particular training event in order to move on to the next phase of training. If they achieve a passing score, they have met the standard. In the military, the term standard is also used broadly to refer to individual and unit performance levels necessary to ensure mission success. To be valid, a selection standard for entering and continuing in an occupation will be correlated with this broader concept of performance. A valid selection standard should not result in a "lowering of military standards." In fact, maintaining military standards is the overarching purpose of validating standards. Because Congress and the public have stressed

[1] In some cases, assigning numerical scores is not intuitive, and subjective scores are necessary. In those cases, the subjective scoring process should be developed by a group of SMEs and tested to ensure that it is applied consistently across raters and rates. Minimum standards on those tests should be established using standard setting panels or direct links between the rating scores and objective measures of performance.

the importance of maintaining standards, this is an important point to highlight. Therefore, extra care should be taken to ensure that any use of the term standard has a clear context.

Occupation-Specific Standards Versus Health and Fitness Standards

An *occupation-specific standard* is a standard used to determine whether an applicant is qualified for a particular job. An example would be a minimum score on a physical test used to determine who is qualified to enter the training pipeline for a physically demanding occupation. Annual fitness tests applied only to members of one occupation are another example. Occupation-specific standards such as these should be tied to concrete occupational requirements. That is, they should exist to help screen out people who are not capable of satisfactory performance in that occupation.

Forcewide health and fitness standards do not serve the same purpose as occupation-specific standards. One goal for health and fitness standards is to establish and maintain a norm or culture of fitness within the overall force. This ensures that members of the force are healthy, which in turn reduces health care costs, injuries on the job, and lost work days to illness and injury. Another goal is to ensure that all personnel are capable of handling physically challenging circumstances that may arise during a mission (e.g., extreme heat). Standards to ensure the health of the force will not be occupation-specific, and they need not be criterion-referenced. Norm-based, gender-specific, and age-specific test scores may instead be preferred. In fact, gender- and age-specific norms are often the best way to evaluate someone's health. For example, research has shown that the amount of body fat that is associated with certain health outcomes differs significantly between men and women, and the range of 5-mile run times for healthy adults changes as we age. For that reason, separate scoring of health measures for men, women, and different age groups may be most appropriate. However, to the extent that the goal is fitness, there may be a need for some criterion-referenced standards to be applied forcewide.

To help clarify when gender-neutrality should be applied and when it should not, the services should document and communicate clearly which of their standards are in place to ensure a generally healthy

force and which are in place to ensure personnel can meet the physical demands of a particular occupation or particular circumstances.

Gender-Neutrality and Bias

In the FY 2014 NDAA, Congress established the legal definition of *gender-neutral standard* in the military:

> The term gender-neutral occupational standard, with respect to a military career designator, means that all members of the Armed Forces serving in or assigned to the military career designator must meet the same performance outcome-based standards for the successful accomplishment of the necessary and required specific tasks associated with the qualifications and duties performed while serving in or assigned to the military career designator. (Pub. L. 113-66, 2013)

By this definition, the concept is very simple and straightforward. If the minimum passing score is the same for women as it is for men, then it is *gender neutral*.

Nevertheless, the term *gender neutral* often can be confusing in practice. For example, some incorrectly deduce that examination of test scores for bias against women or men would not be a gender-neutral activity. That is quite the opposite of what is typically intended when establishing a policy of gender neutrality. In most cases, the intention is not merely that the standard be the same for both genders, but also that the scores on screening tests be equally valid and have the same meaning for both genders—the defining characteristics of an *unbiased standard*. Unbiased standards are standards equally valid in predicting important outcomes for both sexes. Having gender-neutral standards and unbiased standards are both vital for integrating women into combat jobs. Therefore, both should be addressed in the services' efforts.

Bias is probably the least understood concept among policymakers and stakeholders. This is exacerbated by the fact that it can be an emotionally loaded term commonly used by the media and the public in reference to race, gender, or religious discrimination in the workplace. Those uses can be entirely inconsistent with the definitions

that have been adopted by the personnel selection research communities, the Equal Employment Opportunity Commission, and the U.S. Department of Labor. The public often misunderstands bias as occurring whenever two groups score differently on a test; the research community does not define it that way.

In this report, we use *bias* in a very narrow way. The formal and scientific definition of *bias* is "systematic error that differentially affects the performance of different groups of test-takers" (*Standards*, 1999, p. 31). This systematic error is what results in a test being unfair to one group relative to another. In the case of selection and screening tests, we are most concerned with *predictive bias*, which is a type of statistical bias that can take two forms. It can occur when predictive validity differs by group, a phenomenon known as *differential validity*. If the test is a better predictor of performance for one group than it is for another, then the test is considered biased against the group with the lower predictive validity. Or it can occur when the predictive validity is equivalent for both groups but scores underpredict one group's performance relative to another group. For example, a higher score on an entry test involving a physical obstacle course may similarly predict better performance in infantry training for men and women, but the same score on the test for a woman may be predictive of higher performance in training than it is for a man. This test would not have differential validity, but it would be biased, in this example, in favor of men.

Bias is always something that should be examined when there are differences in test scores across groups. In the case of physical testing, gender bias should always be examined, as there are large differences in the average physical capabilities between men and women.[2] Although bias should always be examined in those cases, researchers often discover that no bias exists. A finding of no bias is likely to occur when a test is closely aligned with or valid for predicting on-the-job requirements. For example, in the context of a job that demands that personnel lift 80-pound equipment repeatedly to chest height, we would likely find that a test evaluating this requirement will likely predict

[2] Race and ethnicity differences may also exist. Examination of bias against racial or ethnic groups is also a worthwhile endeavor.

success equally for men and women. That is, even if very few women can meet the 80-pound lift standard on the test and nearly all men can, the test would not be biased against women if it predicts accurately whether they can do that task on the job. Assuming that part of the job is necessary, then establishing that standard on the test would be fair.

Bias is important not only for ensuring fairness of selection practices, but also for ensuring accuracy. A test that is biased against women, for example, is a test that does not do a good job of determining who in that population will be successful on the job. The goal of any selection process in the military should be to measure the qualifications of all personnel and match their KSAs to the job as accurately as possible. Mistakes in selection, even if they occur only for one gender, do not meet that goal.

Validation of Selection Practices

Validation is the process of measuring, quantifying, and collecting evidence to support the use of the test as a selection tool. In other words, if a test is used to identify who is and is not qualified to do the job, then there should be a positive and sufficiently strong relationship[3] between test scores and performance on the job. Higher scores on the test should be associated with better performance.

Job analysis serves as the foundation of all selection validation efforts and should be the first step in establishing validation evidence

[3] The meaning of *sufficiently strong* when referring to correlations in personnel validation can be open to interpretation and should differ depending on the context. For example, correlations as low as 0.10 to 0.30 in some studies could be sufficient to suggest a test is a valid predictor of performance, particularly when the performance predicted is highly complex (that is, it depends on many factors, not just the one measured by the test, such as successfully capturing an enemy target on a mission) and/or the performance data were collected decades after the selection tool information. In those cases, the predictor should be shown to be an equivalent or better predictor of that performance than other reasonable alternatives, or it should be shown to add additional unique information to the prediction. In other cases, as in simulation studies, where the data are collected at the same general time (for example, over a few days) and the performance being predicted in the simulation shares a high degree of content overlap with the predictor (for example, lifting weights as the test and lifting boxes as the performance that is simulated), researchers might expect much higher correlations (0.70 to 0.90) before one would say the relationship is sufficiently strong to justify the test's use.

to support a test's use. Job analysis (also called occupational analysis, task analysis, or work analysis) is the process of establishing an accurate accounting of the tasks or activities that take place in a job. The job analysis should include sufficient detail about the job tasks and activities to determine the physical capabilities required to perform them. Although all validation efforts should be grounded in job analysis, there are several different types of validation that can support a test's use for selection. The two most applicable for physical screening are content validation and criterion-related validation.

Content validation is the process of establishing the degree to which a test adequately captures the entire performance domain of interest. Data requirements usually include judgments from SMEs who are familiar with the test components and the job requirements. If there is a high level of overlap between the test content and on-the-job requirements, the test has high content validity. Content validity is often confused with face validity. *Face validity* is the lay perceptions of a test's validity. If test-takers, instructors, policymakers, and others believe the test seems like it is important for the job, then it has face validity. Face validity is not an acceptable form of validation evidence to support a test's use. Judgments about a test's relevance for the job need to be collected in a systematic manner and supported with concrete evidence to qualify as content validation evidence.

Criterion-related validation is the process of collecting evidence that test scores are correlated with measures of important organizational outcomes. Data requirements usually include test scores from incumbents (i.e., operators) or applicants and measures of performance (e.g., training performance, job performance, errors). There are two types of criterion-related validity: predictive and concurrent. *Predictive validation* involves evidence collected as longitudinal data, i.e., data collected at two different times. Predictor information (data on the selection tests) is collected on personnel at time of hiring, and outcome measures are collected after personnel have been on the job for some period of time. *Concurrent validation* uses evidence from predictors and outcomes data collected around the same time period. It typically involves collecting information about the outcomes of interest (e.g., injuries, job performance) on job incumbents (i.e., operators) and

administering the selection tests to those same incumbents. A *simulation study* is a modified form of a concurrent validation study that may be justified when collecting predictive validation and/or concurrent validation data is not feasible. In a simulation study, participants are measured on a predictor test, trained on how to perform key job activities, and tested on a series of simulations of those activities. If a relationship is shown between the test and the simulated outcomes, and if job analysis data and content analysis of the simulation support the simulation's overlap with key elements of the job, the findings would qualify as reasonable criterion-related validation evidence.

Validation is a complex effort that requires a sound research methodology. This is one reason that validation efforts should be clearly documented. Such documentation allows independent review of the validation effort. Another reason for documentation is that information not documented can get lost over time. If the work is not documented, then as researchers leave, retire, or forget what was done, the institutional knowledge of it deteriorates. Additionally, when inquiries are made by outside parties as to the work that supports the use of current standards, documentation can be easily and quickly provided. Lastly, documenting the work forces the researchers to be clear about their purpose, goals, and procedures, and it illuminates how key terms and issues were interpreted by the researchers and clarifies the limitations of the findings. For all of these reasons, it will be important that the services be asked to thoroughly document their validation efforts.

Summary

The following are among the key points discussed in this appendix:

- The most definitive definitions are set forth in the guidance provided by the personnel research community, and the guidelines provided by Society for Industrial and Organizational Psychology and the American Psychological Association are the source of best practice in establishing job requirements.

- There can be many physical screening points over a career, and these should all involve validated standards. There should be a complete inventory of the measures, tests, evaluations, and other events (broadly defined) that result in someone being excluded because of inadequate physical performance from the occupation before, during, and at the end of training, and across a career. It will be important to recognize this and be clear about which aspects of selection have been validated. The service efforts we tracked focus on entry-level occupational-classification standards. As the first women move through the pipeline in the newly opened occupations, additional effort will be needed to ensure that standards applied at later stages of their careers are also validated.
- Gender-neutral standards are standards that are the same for men and women. However, gender-neutral standards should also be validated and shown to be unbiased (or fair)—i.e., exhibiting the same relationship to job performance for men and women and not underpredicting performance for men or women.
- Documentation of the current work to develop occupation-specific, gender-neutral physical standards should use standard terminology to ensure consistent understanding.

APPENDIX B

Physically Demanding Occupations Open to Women Before 2016

This appepndix focuses on the services' efforts to establish valid physical standards for occupations open to women before 2016. All of the services have physically demanding jobs that fall into this category; however, two of the services (the Navy and the Air Force) have clearly designated a set of occupations they have identified as physically demanding, and both have established a standardized physical screening process for those occupations. Therefore, we discuss the Air Force's and the Navy's screening criteria and their efforts to validate those criteria in detail later in this appendix. First, we provide a brief overview of the status of all of the services' efforts in this area.

The information contained in this appendix comes from our cursory review of published documentation on the existing selection processes for the open occupations, as well as from interviews with representatives from each of the services and the unpublished documentation on the screening processes that they provided to us.

Overview of the Services' Efforts to Establish Physical Standards for Open Occupations

In the Marine Corps and the Army, although there are many occupations known to have physical demands, occupation-specific screening processes to exclude individuals who are not capable of meeting those demands are generally ad hoc, if they exist at all. That is not to say that

many of those demands have not been formally documented or identified as part of the requirements of the job. In fact, throughout all of the Army's MOS job descriptions, the physical demands of the job are explicitly named. For example, the description of the military police occupation (MOS 31B) names the following physical job requirements (Headquarters, Department of the Army, 2015):

- Occasionally lifts 84 pounds, 3 feet and carries 84 pounds, 6 feet as part of a two-soldier team (prorated at 42 pounds per soldier)
- Frequently lifts 42 pounds over head
- Occasionally walks slowly for two out of six hours while carrying 170.9 pounds
- Frequently stands for extended periods of time.

Each MOS is also assigned a physical demands rating according to the following scale (Hollander, Bell, and Sharp, 2008):

- Light: Lift, on an occasional basis, a maximum of 20 pounds with frequent or constant lifting of 10 pounds
- Medium: Lift, on an occasional basis, a maximum of 50 pounds with frequent or constant lifting of 25 pounds
- Moderately Heavy: Lift, on an occasional basis, a maximum of 80 pounds with frequent or constant lifting of 40 pounds
- Heavy: Lift, on an occasional basis, a maximum of 100 pounds with frequent or constant lifting of 50 pounds
- Very Heavy: Lift, on an occasional basis, over 100 pounds with frequent or constant lifting in excess of 50 pounds.

For example, the physical demand rating for military police is listed in Army Pamphlet 611-21 (Section 10-31B) as "heavy lift."

This is just one illustration of an open occupation known to be physically demanding but for which there is no systematic PST in place to determine who should be eligible to be considered for the occupation. Many more such examples can be found across the Army and Marine Corps MOSs. Even in the Navy, where there are only three open occupations that have such formal screening tests in place (discussed later in this appendix), it is likely that there are many more jobs

that also have physical demands that do not use any formal process of physical screening.

When jobs have no screening on physical abilities, the screening process is by definition gender neutral. In other words, the same entry and continuation standards (i.e., no standards at all) are being applied equally to both genders. As a result, the services may presume that having no standards precludes the need for validation of physical standards.

However, it is worth noting that having no standards is not necessarily the best approach to ensuring that personnel in a given job are capable of meeting the physical demands of that job. Instead, the services run the risk that some personnel would be considered unsatisfactory performers, resulting in less than ideal outcomes (e.g., wasted training dollars, wasted personnel resources, and in the worst cases even mission failure or harm to the individual or others). This failure to identify and screen out those who lack the capability to perform is of greater concern when the physical demands of the job are high, and when the frequency and criticality of those duties also are high. For this reason, the Marine Corps, Army, and Navy may want to consider developing a formal set of occupation-specific screening criteria for this broader set of jobs. The Air Force already has that process under way.

The Air Force's Physically Demanding Occupations

The Air Force is the only service that administers a physical aptitude test (called the SAT) to all enlisted personnel upon entry to the service. This test, discussed briefly in Chapter Four, is administered at the MEPS and used to qualify enlisted applicants for certain AFS. Enlisted candidates must complete a lift of at least 40 pounds on the SAT to even qualify to join the Air Force, and many of the AFS in the Air Force do not have any further strength requirement. However, many other jobs do. For those jobs, the lift requirements vary in increments of 10 pounds (i.e., some require a 50, others require a 60, and so on, with a handful requiring scores as high as 90 or 100). The bulk of the

jobs with additional SAT requirements are set at a score of 70. For a complete list of the SAT requirements by AFS, see the AFECD (2013).

Many of the minimum scores for the SAT were established when it was originally instituted in the 1980s. The career field managers adjust job-specific minimum scores only in response to a direct request for reevaluation. The process for adjusting the scores involves researchers visiting site locations to observe and collect data on the physical task requirements (including the types of movements involved in the tasks and the weights of the objects associated with the activities). Those data are then fed into a fixed formula to determine the appropriate minimum SAT score. This formula was established in the 1980s when the SAT was instituted based on simulation-based, criterion-related data analyses; however, the documentation on exactly how the formula was established from those data is scarce.[1]

In implementing its commitment to ensure all jobs have valid, gender-neutral physical standards in place, the Air Force has initiated an effort to establish entirely new SAT minimum scores for all physically demanding jobs. Included in that research effort is the exploration of other physical aptitude measures for use either in addition to or as a replacement for the SAT. The work is sponsored by HAF/A1. The job analysis work to identify and define the physically demanding tasks in each occupation (including surveys of job incumbents and other data collection methods) has been contracted out to RAND, and the validation work (using a simulation-based criterion-related validation approach) and the effort to set minimums on the selection tests have been contracted to an outside consultant. The Air Force anticipates first setting new minimums on the SAT based on this work. After those minimums are set, it will consider the use of any new tests recommended to improve screening as a result of that study.

Although the Air Force has a comprehensive effort under way to identify physical demands and set screening criteria for all enlisted jobs, we are not aware of any effort in place to examine the physical demands of the already-open officer positions.

---

[1] For more background on the SAT, how it was developed, and how minimums have been established and revised over the years, see Sims et al. (2014).

The Navy's Physically Demanding Occupations

Two of the Navy's Warrior Challenge occupations (the SEALs and SWCCs) were discussed in an earlier chapter. The remaining three Challenge occupations—EODs, NDs, and AIRRs—were open to women before 2016 and are discussed here. All three are considered physically demanding by the Navy and have physical screening requirements for training eligibility, physically demanding training elements, and high training attrition rates. The selection process for entry into training in these jobs is also highly competitive, and physical test scores play a large role in who is selected for training.

Minimum eligibility requirements and the training progression for each occupation are summarized in Figures B.1 through B.3. Minimum PST scores for each occupation are shown in Table B.1. Although there are clearly defined minimum scores, each career field also has identified necessary ideal scores for an applicant to be considered competitive. The ideal scores are much more stringent than the minimums. For example, for EOD and ND, the 500-yard swim and the 1½-mile run have a maximum time of 12½ minutes but optimum times of nine minutes and 9½ minutes, respectively. Similarly, the minimum for push-ups and sit-ups is 50, but the optimums are 90 and 85, respectively. Similar optimum scores are needed for AIRR candidates to be competitive.

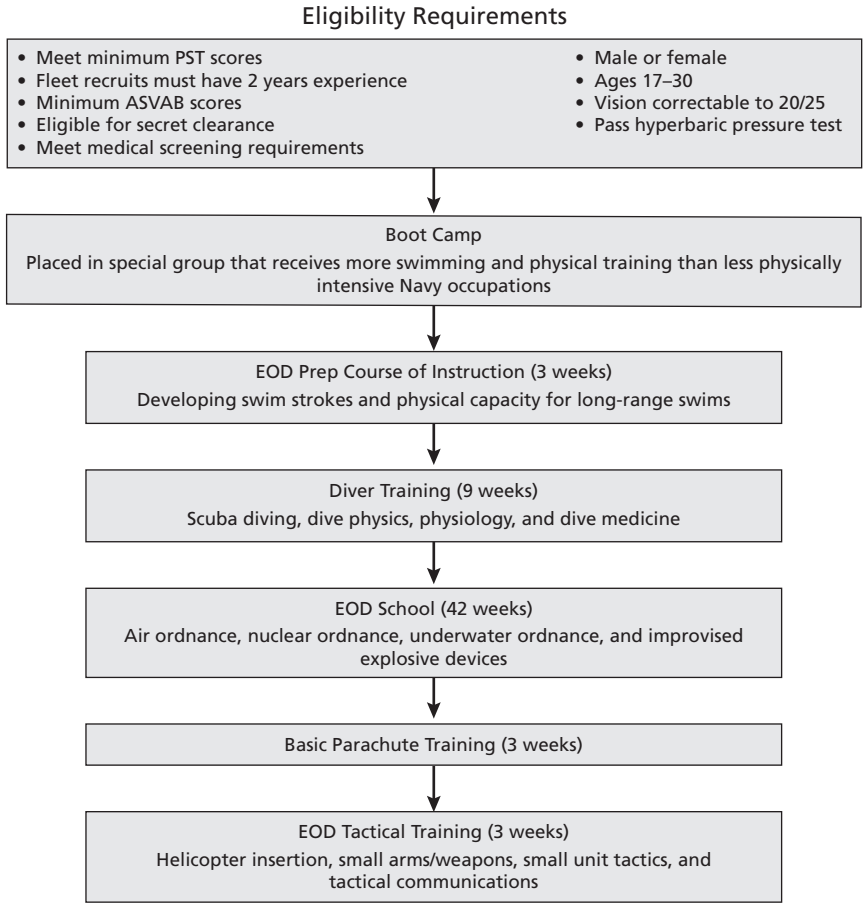
Table B.1
Minimum PST Scores

	EOD Technician	Diver	AIRR
Swim 500 yards, breaststroke or sidestroke (in minutes)	12:30	12:30	12:00[a]
Push-ups (in 2 minutes)	50	50	42
Sit-ups (in 2 minutes)	50	50	50
Pull-ups (in 2 minutes)	6	6	4
Run 1½ miles (in minutes)	12:30	12:30	12:00

SOURCES: Navy Recruiting Command, undated.

[a] AIRR may use sidestroke, breaststroke, American crawl/freestyle, or a combination of all.

Figure B.1
Eligibility and Training Path for Navy EOD Technicians

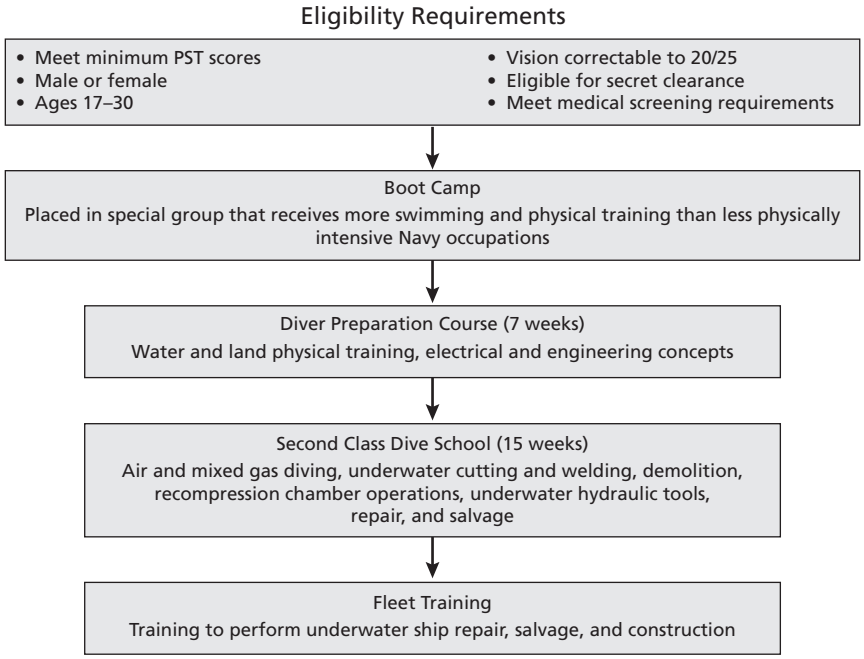


SOURCE: U.S. Navy, undated.
RAND *RR1340/2-B.1*

Enlisted Screening and Assignment

The enlisted occupational assignment and screening process is the same for EOD and ND recruits, and it includes both accession from MEPS stations (the street) and current sailors (the fleet). The majority of accessions come from the street, and the balance come from the fleet.

Figure B.2
Eligibility and Training Path for Navy Divers

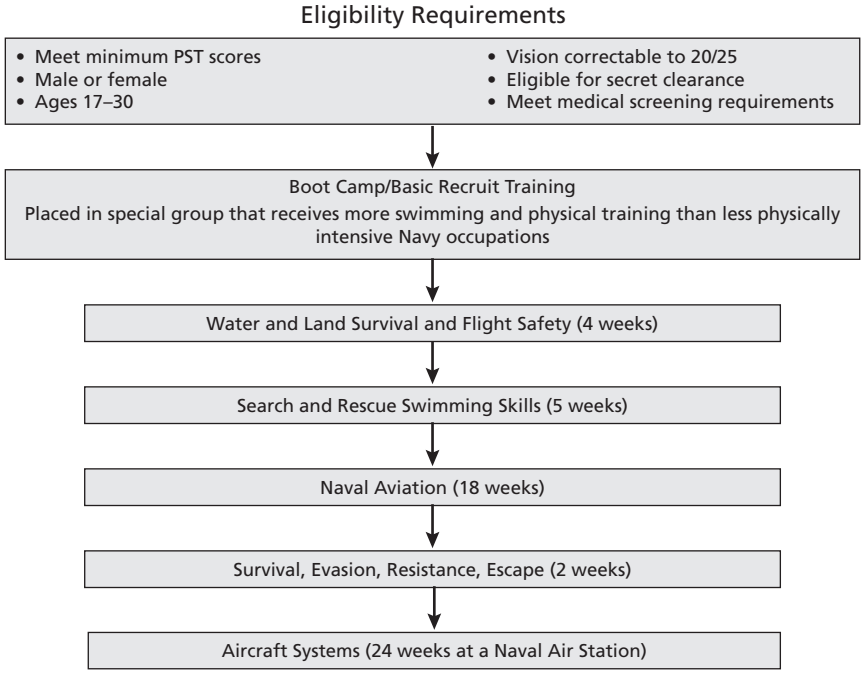


SOURCE: U.S. Navy, undated.
RAND *RR1340/2-B.2*

Street accessions are under the jurisdiction of the Naval Recruiting Command until they graduate from Dive School. Recruits enter their local MEPS, demonstrate an interest in pursuing an EOD or ND career, and take their PST and interview. A committee of evaluators oversees drafting, but PST scores are the strongest determinant of acceptance as an EOD or ND. The career field representatives we interviewed indicated that street recruits are typically of lower overall quality than fleet recruits, such that 75 percent of street recruits will not complete the training requirements necessary to begin EOD and ND operational work.

The number of street accessions who drop out of training determines the number of fleet accessions, who are typically of much higher quality so that pass rates are much higher among fleet than among

Figure B.3
Eligibility and Training Path for Navy AIRRs



SOURCE: U.S. Navy, undated.
RAND *RR1340/2-B.3*

street recruits. In a year with minimal street attrition there would be no fleet accessions. The EOD and ND commanders select which fleet members to send to training from among the five to ten fleet application packages they receive per week.

For both street and fleet applicants, the decision to send an individual to training is determined by the overall quality of his or her application package. The minimum PST scores were raised two years ago based on attrition data collected at the Center for EOD and Diving, which showed the cutoff point over which no recruits scoring at that level completed training.

Typically, accepted individuals demonstrate sub–nine-minute swims and runs, 20 or more pull-ups, and 100 push-ups and sit-ups—all scores that are considerably better than the minimum standard.

In our interviews, the EOD career field representative reported that the median PST scores for 33 recent EOD officer accessions (from a pool of 100 candidates), including men and women, were swim: 8:37 minutes; push-ups: 103; sit-ups: 95; pull-ups: 16; run: 9:33 minutes. Per the AIRR community manager, many male applicants complete their swim in 6:30 minutes, run in 8:00 minutes, 100 push-ups, and 10 pull-ups.

Therefore, according to EOD and ND community leaders, the Navy could decide to raise the minimum standard again and the increase would not have a meaningful effect on the quality of the applicants or of those who are chosen to attend training—the minimums are not actually driving those pools. EOD and ND community leaders indicated that even "if all the standards were increased by one-third, we would still get the same number and quality of packages."

Many aspects of the AIRR screening and occupational assignment process are the same as for EODs and NDs, including the acceptance of street and fleet accessions, and inclusion of AIRR, EOD, and ND recruits in the Basic Recruit Training's 800 Company. In particular, demonstration of the minimum PST standards typically is not sufficient to warrant a contract and invitation to enter training for AIRR recruits.

A difference between the EOD/ND process and the AIRR process is that AIRR recruiters and community leaders use an "auto-qualification" formula (autoqual). The formula determines which recruits qualify automatically based on a combination of their PST component and overall scores, ASVAB score, vision, region, recruiting district, age, height, and whether they have waivers, legal issues, and a desired program. The autoqual cutoff score is 1,750, though we were told that waivers can be given even if that number is not reached. The history behind the autoqual and the actual formula is unknown to us.[2]

The fraction of applicants that meet the autoqual threshold varies by month and recruitment district. Recruiters prefer to keep all street recruits in the delayed entry pipeline for about six months before send-

[2] We requested and received a copy of the autoqual spreadsheet, but the actual formula is locked—we can see only the various input fields.

ing them off to recruit train. During those six months, recruits must successfully complete a PST every 45 days and then another 15 or fewer days prior to departure.

In recent years, the draft goal has included recruitment of 20 women annually, or just less than 10 percent of the total draft, and 242 men. The goal of 242 is higher than a prior goal of 192; it was raised because of high training attrition over the past few years.

Recruiters typically reach the 20-female quota, but not easily because few women meet the minimum PST requirements, particularly the four pull-ups. That said, the 20 who fulfill the selection requirements always are very competitive and typically include three to four autoqual women. Anecdotally, according to the AIRR community manager, women are not discouraged by the pull-up requirement. They are more concerned about the actual requirements on the job, such as pulling people through the water and otherwise handling the water challenges.

All qualifying applicants enter a draft that goes to the AIRR community office, which then selects from the pool to meet the number needed (the yearly goal is 262). As the application packets come in, the community manager reviews them and picks the strongest candidates, making his final selections based largely on his judgment. When he does a draft, he focuses more on the swim and run times than on the total score; if a candidate has high strength numbers but a low swim time, he or she "might not be the best fit for this occupation." Yet candidates need strong strength scores as well to be able to complete the expected push-ups and pull-ups in boot camp.

During training, beginning at rescue swimmer school, recruits must pass the Swimmer Fitness Test (SFT). Sports physicians designed the SFT in 2001–2002 to include pull-ups, a 1-mile walk carrying a 50-pound dumbbell, a 500-meter freestyle swim with gear on, and a two-person buddy swim for another 400 meters. All events are timed, with timed rests between events. The events are intended as simulations of occupational requirements. SFT is required once a year to maintain the AIRR occupational qualification, though individuals often end up doing it every quarter in addition to the standard Navy PST.

The female screening process for the three occupations is exactly the same as that for men, including the same PST requirements and training. Each occupation has some successful women on the job, but the numbers are generally still small. For example, there are currently 25 female AIRRs, who entered the occupation from graduating classes beginning in 1990 and through 2014. Most years, only one to two female AIRRs joined from each graduating class, though in 2011 and 2012 there were seven and six, respectively. Since 2010, as we show in Table B.2, the female recruiting goal has been 20, with five or fewer women completing training and graduating to the fleet.

Officer Screening and Assignment

For all three occupations, officer applicants come from ROTC, officer candidate school, and the Naval Academy. Screening takes place in fall and spring of Naval Academy and ROTC junior year, with an evaluation process the following summer. The evaluating unit ranks applicants based on physical and academic performance, and the rankings are presented to a formal accession board that includes non-Navy lead-

Table B.2
Female AIRR Recruiting Goal

Graduation Year	Goal	Shipped to Recruit Training	Graduates
2008	57	20	1
2009	15	15	2
2010	20	21	2
2011	20	19	5
2012	20	15	5
2013	20	20	2
2014	20	18	N/A
2015	20	9	N/A

NOTE: N/A indicates that not enough time had passed for all of the students to have the opportunity to graduate. As a result, final numbers of graduates were not known.

ers. For EOD, the board typically selects 27 to 28 individuals to enter officer training, and the rest enter from the enlisted EOD occupation. As of December 31, 2014, the EOD officer group included 417 men and 12 women. The female officers all scored above the average male applicants, including on pull-ups (anecdotally, the women who were chosen each completed well over 15 pull-ups).

Training

Male street accessions to EOD, ND, and AIRR all enter Basic Recruit Training as part of the 800 Company, in which they take part in considerably more physical training and swimming than sailors entering less physically intensive occupational tracks. All EOD, ND, and AIRR women are part of a separate female division equivalent to the 800 Company, although they complete PT with the men.

Additional trainings for both fleet and street accessions are designed to prepare recruits for on-the-job physical demands. To start with, all recruits—both male and female—learn to operate wearing heavy equipment. For example, the EOD bomb suit itself weighs 80 pounds and is loaded with additional weight from parachutes, body armor, and 20 pounds of demolition equipment. EODs must be able to carry the weight of that equipment in an operating environment. Additionally, with a combat load and backpack (120 to 150 pounds in total), just to get in a vehicle is very physically demanding. In-water proficiency drills are also designed to replicate skills and physical strength necessary on the job. Recruits learn to surface under rough conditions and inflate a buoyancy compensator while in full equipment and treading water.

In AIRR, recruits enter Air Division School, followed by Rescue Swimmer School, and then move on to either AWR (tactical) or AWS (nontactical) A-School.[3] Selection into AWR or AWS depends on fleet needs and is determined from the top down. Top recruit performers are typically permitted to choose between the two, and the rest of the recruits are assigned to go where needed.

[3] Women enter Air Division School with the February and June cohorts only.

Though some data provided to us in our interviews show that PST times are useful predictors of training completion for these occupations,[4] EOD and Diver trainers' holistic impression is that level of underwater comfort during training is also a strong predictor of whether a recruit will complete or drop out of training. Typically, there is about 75 percent of attrition from training among the street recruits, less among fleet recruits. Most of the ND attrition takes place during the 21-day prep course.

Attrition in these occupations tends to be high, but it varies across the different training and selection steps. For example, AIRR attrition from boot camp is around 12 percent, and attrition from Rescue Swimmer School has been around 40 percent to 45 percent in the past few years. According to our interviewees, recruits attrit from Rescue Swimmer School for a number of reasons, including that the job is not what they thought it would be, they cannot accomplish the physical requirements in the pool with full equipment, and they are not comfortable in the water. There is very little attrition after Rescue Swimmer School, though there are still a few (around three to four a year), usually for behavioral issues. AIRR also accepts BUD/S dropouts into training, and while they are good physical candidates, they tend to still have a high attrition rate, mostly because they did not really want to pursue the AIRR occupation.

Navy's Process for Validating the EOD, ND, and AIRR PST Standards

According to ND and EOD community managers, the PST screening standards have not been significantly modified in more than 30 years. No information could be provided on how the original PST tests were selected or why; however, a few published studies have explored the criterion-related validity of the test for NDs and EODs. Those studies generally have not found strong support for using the PST elements to screen personnel for entry into these occupations (for examples, see

[4] A scatter plot of 714 EOD and ND prep students provided by our interviewees shows that on the combined run and swim time (in seconds) taken at the end of Basic Recruit Training, only 11 percent (37 of 333) prep graduates took longer than 1,300 seconds (i.e., 21 minutes and 40 seconds), compared with 34 percent (129 of 381) of those who dropped out.

Marcinik, Hyde, and Taylor, 1993; and Hodgdon et al., 1998). However, there were limitations to those studies, including that the trainee participants had already been screened on the PST; therefore, their physical aptitude represented an already restricted range of scores. As a result, more research on the use of the PST or other physical aptitude tests for screening personnel in these occupations would certainly be warranted.

Training content is reviewed and updated regularly. Naval Education and Training Command (NETC) procedures require a review of training once every three to five years through a process called the "human performance requirements review." Detailed instructions for how to conduct the review are documented in official Navy policy.

The review process includes SMEs' review of a given course of instruction. Particular documented triggers will lead to a required resubmission of the training plan. For example, if the SMEs determine that the training does not cover an important area and/or it requires additional resources to do so (like more days), the training plan will then go up the chain of command for approval. In addition, if, during a given course review, NETC discovers a high attrition rate, it would then speak with course trainers to learn whether there are specific activities that considerable numbers of trainees are not able to accomplish. NETC will then review the occupational need for that particular performance requirement. Navy Manpower Analysis Center sets the original occupational standards,[5] so it also plays a role in the revision of the given training component. In addition, following NAVEDTRA 135C, NETC conducts an annual formal review on every course, including test item analysis, student critiques, attrition, etc. NAVEDTRA 135C specifies which organizations should be included in each course review; inclusion varies by the course of instruction. This review leads to an immediate change when there is a safety issue. A safety risk team is involved regarding training in a high-risk course. Attrition is not typically broken down by gender.

[5] Navy Personnel Command (NAVPERS) 18068F Volume I (2016a) lists occupational standards, and Volume II (2016b) lists Navy occupational codes.

References

"Air Force Special Tactics Fitness Training," webpage, Military.com, June 1, 2011. As of March 21, 2018:
https://www.military.com/military-fitness/air-force-special-operations/air-force-special-tactics-fitness-training

Aleton, Sophi, Zoe Cohen, Jamey Cummings, and Michael Gray, "So, You Want to Be a Frogman?: Determining What It Takes to Become a U.S. Navy Seal," unpublished academic paper, December 4, 2002.

Aspin, Les, "Direct Ground Combat Definition and Assignment Rule," memorandum to the Secretary of the Army, Secretary of the Navy, Secretary of the Air Force, Chair of the Joint Chiefs of Staff, Assistant Secretary of Defense for Personnel and Readiness, and Assistant Secretary of Defense for Reserve Affairs, Washington, D.C., January 13, 1994.

Bretton, Gene, and Linda Doherty, *Initial Recommendations for Hell Week*, Navy Personnel Research and Development Center, San Diego, Calif., January 31, 1979.

Burrelli, David F., *Women in Combat: Issues for Congress*, Washington, D.C.: Congressional Research Service, May 9, 2013. As of May 23, 2016:
https://fas.org/sgp/crs/natsec/R42075.pdf

Charlton, Samuel G., "Measurement of Cognitive States in Test and Evaluation," in Samuel G. Charlton and Thomas G. O'Brien, eds., *Handbook of Human Factors Testing and Evaluation*, Mahwah, N.J.: Lawrence Erlbaum Associates, 2002.

Dawes, Robyn M., and Bernard Corrigan, "Linear Models in Decision Making," *Psychological Bulletin,* Vol. 81, No. 2, February 1974, pp. 95–106.

Doherty, Linda M., Thomas Trent, and Gene E. Bretton, *Counterattrition in Basic Underwater Demolition/SEAL Program: Selection and Training*, No. NPRDC-SR-81-13, Navy Personnel Research and Development Center, San Diego, Calif., 1981.

Donley, Michael B., "Air Force Implementation Plan for Integrating Women into Career Fields Engaged in Direct Combat," memorandum from the Secretary of the Air Force to the Secretary of Defense, Washington, D.C., April 24, 2013.

Hardison, Chaitra M., Carra S. Sims, and Eunice C. Wong, *The Air Force Officer Qualifying Test: Validity, Fairness, and Bias*, Santa Monica, Calif.: RAND Corporation, TR-744-AF, 2010. As of May 03, 2016: http://www.rand.org/pubs/technical_reports/TR744.html

Hardison, Chaitra M., Susan D. Hosek, and Chloe E. Bird, *Establishing Gender-Neutral Physical Standards for Ground Combat Occupations, Vol. 1: A Review of Best-Practice Methods*, Santa Monica, Calif.: RAND Corporation, RR-1340/1, 2018.

Headquarters, Air Education and Training Command, A3T, homepage, undated. As of May 2, 2018: http://www.aetc.af.mil/

Headquarters, Department of the Army, "Military Occupational Classification and Structure," Pamphlet 611-21, undated. As of May 24, 2016: https://www.milsuite.mil/book/groups/smartbookdapam611-21

Hodgdon, James A., Marcie B. Beckett, Tracy Sopchick, W. Keith Prusaczyk, and Harold W. Goforth, Jr., *Physical Fitness Requirements for Explosive Ordnance Disposal Divers*, No. NHRC-98-31, San Diego, Calif.: Naval Health Research Center, 1998.

Hoffmann, Dick, *BUD/S Attrition: A Review of Past Research and Current Practices*, paper presented to Dr. Myron Dembo in fulfilment of course Educational Psychology and Technology 614, Social Psychology of Education, University of Southern California, unpublished, November 18, 2002.

Hollander, Ilyssa E., Nicole S. Bell, and Marilyn Sharp, *Physical Demands of Army Military Occupational Specialties: Constructing and Applying a Crosswalk to Evaluate the Relationship Between Occupational Physical Demands and Hospitalizations*, No. USARIEM-TR-T08-06, Boston, Mass.: Social Sectors Development Strategies, Inc., 2008.

Joint Committee on Standards for Educational and Psychological Testing, *Standards for Educational and Psychological Testing*, Washington, D.C.: American Educational Research Association, 2014.

Lamothe, Dan, "First Army Ranger School with Women Opens with 16 Passing Initial Test," *Washington Post*, April 20, 2014. As of May 17, 2016: http://www.washingtonpost.com/news/checkpoint/wp/2015/04/20/first-army-ranger-school-with-women-opens-with-16-passing-initial-test/

Mabus, Ray, "Marine Corps Women in the Service Review Implementation Plan," memorandum from the Secretary of the Navy to the Secretary of Defense, Washington, D.C., May 2, 2013a.

———, "Department of the Navy Women in the Service Review Implementation Plan," memorandum from the Secretary of the Navy to the Secretary of Defense, Washington, D.C., May 2, 2013b.

Manacapilli, Thomas, Carl F. Matthies, Louis W. Miller, Paul Howe, P. J. Perez, Chaitra M. Hardison, H. G. Massey, Jerald Greenberg, Christopher Beighley, and Carra S. Sims, *Reducing Attrition in Selected Air Force Training Pipelines*, Santa Monica, Calif.: RAND Corporation, TR-955, 2012. As of May 24, 2016: http://www.rand.org/pubs/technical_reports/TR955.html

Marcinik, Edward J., Dale E. Hyde, and W. Fred Taylor, *Validation of the U.S. Navy Fleet Diver Physical Screening Test*, No. NMRI-93-79, Bethesda, Md.: Naval Medical Research Institute, 1993.

Marine Corps Operational Test and Evaluation Activity, *Ground Combat Element Integrated Task Force: Experimental Assessment Plan*, Washington, D.C.: Headquarters Marine Corps, May 2014.

Marine Corps Forces Special Operations Command, "Selection and Training," webpage, undated. As of March 20, 2018: https://marsoc.com/selection-and-training

Marine Corps Forces Special Operations Command, Marine Raider Training Center, "Assessment and Selection Program (A&S)," webpage, undated. As of March 20, 2018: http://www.marsoc.marines.mil/Units/Marine-Raider-Training-Center/Assessment-Screening/

MARSOC—*See* Marine Corps Forces Special Operations Command.

McDonald, D. G., J. P. Norton, and J. A. Hodgdon, "Training Success in U.S. Navy Special Forces," *Aviation, Space, and Environmental Medicine*, Vol. 61, No. 6, 1990.

McHugh, John M., "Plan for Integration of Female Leaders and Soldiers Based on the Elimination of the 1994 Direct Ground Combat Definition and Assignment Rule (DGCAR)," memorandum from Chairman of the Joint Chiefs of Staff to the Secretary of Defense, Washington, D.C., April 19, 2013.

McRaven, William H., "U.S. Special Operations Command Implementation Plan for Elimination of Direct Combat Assignment Rule," memorandum from Admiral, U.S. Navy, Commander, to Chief of Staff, U.S. Army; Commandant, U.S. Marine Corps; and Chief of Naval Operations; Chief of Staff, U.S. Air Force, Washington, D.C., March 22, 2013.

MCOTEA—*See* Marine Corps Operational Test and Evaluation Activity.

Mills, Lisa J., and Janet D. Held, "Optimizing U.S. Navy Seal Selection," Navy Selection & Classification Office, Arlington, Va., unpublished, undated.

Mills, Lisa J., and Sean Robson, "Individual Characteristics Related to SEAL Training Success," Navy Selection, Classification & Surveys (N141), Radford University, unpublished briefing, undated.

Naval Education and Training Command, *Navy School Management Manual*, NAVEDTRA 135C, Pensacola, Fla., March 2010.

Naval Health Research Center and U.S. Marine Corps Training and Education Command, *Analysis in Support of the Women in Service Restriction Review Study*, undated. As of May 25, 2016:
http://www.defense.gov/Portals/1/Documents/wisr-studies/USMC%20-%20NHRC%20Analysis%20in%20Support%20of%20the%20Women%20in%20Service%20Restrictions%20Review%20Study.pdf

Naval Military Personnel Manual, SEAL/EOD/SWCC/DIVER/AIRR Physical Screening Testing, Standards and Procedures, MILPERSMAN 1220-410, January 2013. As of March 20, 2018:
http://www.public.navy.mil/bupers-npc/reference/milpersman/1000/1200classification/documents/1220-410.pdf

———, Special Warfare Boat Operator (SB) Rating, MILPERSMAN 1220-400, November 2016. As of March 20, 2018:
https://www.sealswcc.com/pdf/milpersman-1220-400-special-boat.pdf

Naval Special Warfare Center, "Navy SWCC General Requirements," webpage, May 2016. As of March 20, 2018:
https://www.sealswcc.com/navy-swcc-general-requirements.html

NAVPERS—*See* U.S. Navy Personnel Command.

Navy Recruiting Command, "Aviation Rescue Swimmer," webpage, undated. As of May 16, 2016:
http://www.navy.com/careers/special-operations/air-rescue.html#ft-training-&-advancement

Office of Personnel Management, "Assessment & Selection: Other Assessment Methods—Physical Ability Tests," webpage, undated. As of May 17, 2016:
https://www.opm.gov/policy-data-oversight/assessment-and-selection/other-assessment-methods/physical-ability-tests/

———, *Delegated Examining Operations Handbook: Guide for Federal Agency Examining Offices*, 2007. As of March 20, 2018:
http://www.archives.gov/careers/competencies/resources/opm-examining-operations-handbook.pdf

Office of the Under Secretary of Defense for Personnel and Readiness, *Report to Congress on the Review of Laws, Policies and Regulations Restricting the Service of Female Members in the U.S. Armed Forces*, Washington, D.C.: U.S. Department of Defense, February 2012.

OPM—*See* Office of Personnel Management.

Panetta, Leon E., "Statement on Women in Service," statement delivered in Pentagon Press Briefing Room, Washington, D.C., January 24, 2013.

Panetta, Leon, and Martin E. Dempsey, "Elimination of the 1994 Direct Ground Combat Definition and Assignment Rule," memorandum from Secretary of Defense and Chairman, Joint Chiefs of Staff to the U.S. military secretaries, Washington, D.C., January 24, 2013.

Prusaczyk, W. Keith, Jack W. Stuster, Harold W. Goforth Jr., Tracy Sopchick Smith, and L. T. Meyer, *Physical Demands of US Navy Sea-Air-Land (SEAL) Operations*, No. NHRC-95-24, San Diego, Calif.: Naval Health Research Center, 1995.

Prusaczyk, W. Keith, Jack W. Stuster, Harold W. Goforth Jr., Marcie B. Beckett, and James A. Hodgdon, *Survey of Physically Demanding Tasks of US Navy Explosive Ordnance Disposal (EOD) Personnel*, No. NHRC-98-35, San Diego, Calif.: Naval Health Research Center, 1998.

Public Law 103-160, National Defense Authorization Act for Fiscal Year 1994, November 30, 1993.

Public Law 113-66, National Defense Authorization Act for Fiscal Year 2014, December 26, 2013. As of March 21, 2018: https://www.congress.gov/113/plaws/publ66/PLAW-113publ66.pdf

Public Law 113-291, Carl Levin and Howard "Buck" McKeon National Defense Authorization Act for Fiscal Year 2015, December 19, 2014. As of March 21, 2018: https://www.congress.gov/113/plaws/publ291/PLAW-113publ291.pdf

San Diego State University School of Business Administration, "NSW Consulting Report on SEAL Database Analysis," unpublished briefing, 2002.

Sharp, Marilyn, *Development of Military Occupation-Specific Physical Employment Standards*, unpublished Institutional Review Board protocol, February 19, 2014a.

———, *Development of Military Occupation-Specific Physical Employment Standards: Study 3*, unpublished Institutional Review Board protocol, May 28, 2014b.

Society for Industrial and Organizational Psychology, *Principles for the Validation and Use of Personnel Selection Procedures*, 4th ed., Bowling Green, Ohio, 2003.

TECOM—*See* Training and Education Command.

Thomas Group, "Macro Assessment Outbrief," presentation to Joseph Maguire, Commander, Naval Special Warfare Command, in unpublished briefing provided to RAND, August 10, 2006.

Training and Education Command, Marine Air-Ground Task Force Training and Education Standards Division, *Analysis and Assessment of MOS Physical Performance Standards*, briefing to authors, December 2013.

Tversky, Amos, and Daniel Kahneman, "Judgment Under Uncertainty: Heuristics and Biases," *Science*, Vol. 185, No. 4157, September 27, 1974, pp. 1124–1131.

U.S. Air Force, "U.S. Air Force Special Ops," webpage, undated. As of March 20, 2018: https://www.airforce.com/careers/featured-careers/special-operations

———, *Air Force Enlisted Classification Directory (AFECD), The Official Guide to the Air Force Enlisted Classification Codes*, Washington, D.C.: Headquarters Air Force Personnel Center, April 30, 2013a.

———, *Air Force Officer Classification Directory (AFOCD), The Official Guide to the Air Force Officer Classification Codes*, Washington, D.C.: Headquarters Air Force Personnel Center, April 30, 2013b.

———, *Classifying Military Personnel (Officer and Enlisted)*, Air Force Instruction 36-2101, June 25, 2013c.

U.S Army, Airborne and Ranger Training Brigade, homepage, undated. As of May 17, 2016:
http://www.benning.army.mil/infantry/rtb/

USASOC—*See* U.S. Army Special Operations Command.

U.S. Army Special Operations Command, "75th Ranger Regiment," website, undated. As of May 17, 2016:
http://www.soc.mil/Rangers/75thRR.html

U.S. Marine Corps, "Being a Marine: Career Roles and Leadership Traits," webpage, undated. As of June 4, 2018:
https://www.marines.com/being-a-marine/life-in-the-corps.html

———, *Training and Readiness Manual Group (TRMG) Charter Terms of Reference*, NAVMC 3500.106, July 2011. As of May 17, 2016:
http://www.marines.mil/Portals/59/Publications/NAVMC%203500.106.pdf

U.S. Navy Personnel Command, *Manual of Navy Enlisted Manpower and Personnel Classifications and Occupational Standards*, Volume I: *Navy Enlisted Occupational Standards*, NAVPERS 18068F, January 2016a.

———, *Manual of Navy Enlisted Manpower and Personnel Classifications and Occupational Standards*, Volume II: *Navy Enlisted Classifications (NEC)*, NAVPERS 18068F, January 2016b.

U.S. Navy Recruiting Command, *Navy Recruiting Manual–Enlisted*, Volume IV–*Programs and Classification*, COMNAVCRUITCOMINST 1130.8J, May 2011. As of March 20, 2018:
http://navybmr.com/study%20material/CNRCINST%201130.8J%20(VOLUME-IV).pdf

U.S. Naval Special Warfare Command, unpublished statistics produced by Michael Caviston, Basic Training Command, 2014.

Vickers, Ross R., James A. Hodgdon, and Marcie B. Beckett, *Physical Ability-Task Performance Models: Assessing the Risk of Omitted Variable Bias*, San Diego, Calif.: Naval Health Research Center, September 4, 2008.