Draft Research Report

Defining Physical Standards for Physically Demanding Jobs

A Review of Methods

Chaitra M. Hardison
Susan D. Hosek
Chloe E. Bird

RAND National Defense Research Institute

RR-1340/1-OSD
December 2015
Prepared for the Office of the Secretary of Defense

This document has not been formally reviewed or edited. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors. RAND® is a registered trademark.

Preface

Since the establishment of the all-volunteer force in 1973, representation of women in the military has increased to 15 percent, and an increasing number of military occupations have been opened to them. On January 24, 2013, the Secretary of Defense (SecDef) announced that the last remaining policy restricting the service of women, the direct ground combat exclusion rule, would be rescinded. Women will be allowed to serve in any occupation and any assignment for which they can meet the occupational standards. The SecDef also directed the military services to validate their occupational standards to ensure that the standards appropriately reflect occupational requirements and are gender neutral.

The research reported here supports the review and development of gender-neutral physical standards for physically demanding occupations in the military. The first phase of the study, documented in this report, identified the best-practice methods for developing physical standards relevant to these military occupations. The second phase of the study reviewed the methods the services are using to validate their occupational standards in response to the Secretary's guidance.

The research is sponsored by the Office of the Under Secretary of Defense for Personnel and Readiness and 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 Marine Corps, the defense agencies, and the defense Intelligence Community.

Contents

Preface ... ii
Figures ... v
Tables ... vi
Summary ... vii
Methodological Approaches to Establishing Physical Job Requirements ... viii
1. Identify Physical Demands ... ix
2. Identify Potential Screening Tests ... ix
3. Validate and Select Tests ... x
4. Establish Minimum Scores ... xi
5. Implement Screening ... xi
6. Confirm Tests Are Working as Intended ... xii
Final Thoughts ... xii
Acknowledgments ... xiii
Abbreviations ... xiv
Chapter One. Introduction ... 1
Why Standards? ... 2
Occupation-Specific Physical Standards and Entry into Military Service ... 5
Study Approach ... 7
Organization of This Report ... 7
Chapter Two. Methodological Approaches to Establishing Physical Job Requirements ... 9
Six-Stage Process ... 10
Chapter Three. Identify the Physical Demands of the Job ... 13
Methods for Conducting a Job Analysis ... 13
Reasons to Conduct a Careful Job Analysis ... 15
Chapter Four. Identify Potential Screening Tests ... 19
Taxonomies of Physical Aptitudes ... 19
Examples of Tests Studied for Use in Employment Settings ... 21
Selecting Candidate Tests ... 21
Chapter Five. Validate and Select Tests ... 25
Construct Deficiency and Construct Irrelevance ... 26
Content Validity ... 28
Criterion-Related Validity ... 30
Convergent and Discriminant Validity ... 33
Fairness: Adverse Impact and Predictive Bias ... 34
Additional Considerations in Collecting Validation Evidence ... 36
Document the Process ... 36
Apply Appropriate Statistical Methods ... 36

Anticipate Potential Weaknesses in a Study's Methodology and Criticisms of the Results 39
Collect Multiple Sources of Evidence 39
Chapter Six. Establish Minimum Scores 41
Use of Job Analysis Data to Set the Minimum Score 42
Use of Expert Panels to Set the Minimum Score 42
Capture Expert Judgments About Minimum Performance on the Job 43
Capture Expert Judgments About Minimum Performance on the Test 43
Methods for Obtaining Expert Judgments for Setting Standards 44
A Well-Designed Study 44
Chapter Seven. Implement Screening 49
When the Test Should Be Administered 49
Standardize the Test Administration Procedures 50
Informing Applicants About the Test 51
Consider Phasing the Test in Gradually 51
Chapter Eight. Confirm that the Tests Are Working as Intended 53
Institutionalize Research to Support Policy Changes 54
Reexamine Job Analyses 54
Continuously Collect Longitudinal Predictive-Validation Evidence 54
Reexamine Test Score Minimums 54
Collect Test-Taker Reactions and Job-Incumbent Perceptions of the Tests 55
Evaluate Whether Administration Procedures Are Being Followed 55
Conduct Additional Research as Needed 55
Ongoing Personnel Research Efforts Are Not New 55
Chapter Nine. Final Thoughts 57
Phase II of the RAND Study 59
Glossary of Terms 61
References 65

Figures

Figure S.1. Six Stages in Developing Physical Standards viii
Figure 1.1. Enlistment and Initial Training of Military Enlisted Personnel 6
Figure 2.1. Six Stages in Developing Physical Standards 11
Figure 5.1. Conceptual Validation Linkages 26
Figure 5.2. Construct Deficiency 27
Figure 5.3. Construct Irrelevance 27

Tables

Table 3.1. Example Considerations in Identifying the Physical Demands of the Job 17
Table 4.1. Fleishman's Physical Ability Domains..... 20
Table 4.2. Examples of Tests Used to Measure Physical Abilities in the Different Domains 21
Table 4.3. Example Considerations in Identifying Potential Screening Tests..... 24
Table 5.1. Example Considerations in Validating and Selecting Tests 37
Table 6.1. Example Considerations in Establishing Minimum Test Scores..... 47
Table 7.1. Example Considerations in Implementing Screening..... 52
Table 8.1. Example Considerations for Confirming that Tests Are Working as Intended..... 53

Summary

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 (SecDef, 1994) and the intention to "integrate women into occupational fields to the maximum extent possible" (U.S. Department of Defense, 2013). The rule restricted assignments of women 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 including physically demanding tasks the "vast majority" of women cannot do (SecDef, 1994). In announcing the decision to eliminate the rule, the Secretary stated:

> 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.

In 2016, previously closed occupations will be opened to women who can meet occupation-specific, gender-neutral standards of performance. The Joint Chiefs of Staff have established key requirements for implementing this policy change that must be met prior to opening the occupations. These include validating performance standards for military occupations, with special attention to those occupations closed to women.

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 addresses two research objectives. The first is 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 will be to review the methodologies being used by the military services to set gender-neutral standards. This report provides the results of work conducted toward the first research objective.

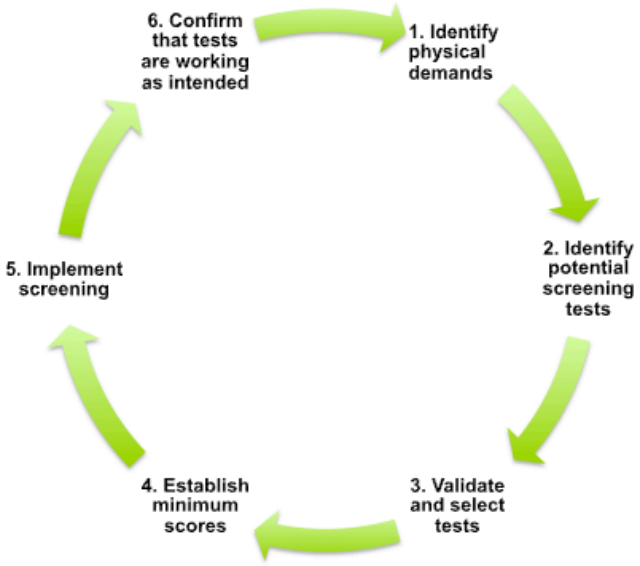
Throughout this report, we use the term *standards* or *physical standards* to refer to occupation-specific criteria that applicants must meet to enter or remain in a particular career field or specialty. We are concerned with standards that are used to make selection decisions—that is, decisions made that may exclude people from entering or continuing in a job. *Gender-neutral standards* are based only on the physical capabilities required to perform the job, are the same for men and women, and should not differentially screen out a higher proportion of members of one gender who are, in fact, able to perform the job. Thus, the challenge for the military services is to identify a set of standards that is the same regardless of gender and valid in predicting job performance for both sexes.

Civilian employers whose jobs are physically demanding have long faced scrutiny regarding the appropriateness and equity of their standards. As DoD embarks on the process of developing gender-neutral physical standards, it can expect similar scrutiny—and, for this reason, wishes to employ appropriate methods in this endeavor. To assist the military services in developing general and occupation-specific standards that are relevant to performance, this interim report describes methods related to physical standard development.

Methodological Approaches to Establishing Physical Job Requirements

The methods for developing physical standards can be organized in six general stages (see Figure S.1). Each element in the process for establishing physical job requirements provides support for the use or exclusion of a set of selection procedures. Carrying out this process requires expertise in a variety of domains, including industrial and organization psychology, exercise physiology or a related field, psychometrics, and statistics. These experts rely on the expertise of subject matter experts 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 steps described here and, importantly, the documentation of the actions taken are critical to developing defensible physical standards.

Figure S.1. Six Stages in Developing Physical Standards



1. Identify Physical Demands

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 and the job analysis will be used to design an appropriate selection system. So the job analysis should identify and describe in detail the physically demanding tasks the applicants would need 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 subject matter experts 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. Similar issues arise in setting standards for continuing in a job, but the test subjects would include 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.

2. Identify Potential Screening Tests

Identifying potential tests that might be used to screen job applicants (or job incumbents) is the next step in developing physical standards. In this context, we use screening to refer to evaluation of individuals' physical skills relevant for performing job tasks. 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 may 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 are 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 the ability to accurately replicate 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. Further, 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 more than one job may be a more feasible approach.

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

3. Validate and Select Tests

The third step 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 is to provide evidence that the selection test predicts important outcomes on the job.

Best practice requires that 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 may be selected who are not capable of performing on the job or candidates may 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 expertise and require a careful design before data collection begins to ensure results are as accurate as possible and avoid bias toward any group of applicants. Finally, organizations should seek multiple sources of validation evidence whenever possible.

4. Establish Minimum Scores

The next step in the process is to establish the minimum scores that will reflect acceptable performance on the job. The goal in this step 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, 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 the Secretary's commitment to not "reducing the qualifications for the job."

The process of establishing minimum cutoff scores, referred to as standard setting, is distinct from validation. When used in 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 may 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.

5. Implement Screening

Once the previous steps have been completed and clear instructions for the proper test administration procedures devised, it is appropriate to begin using 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 that is consistent with the results of the validation and standard-setting efforts.

The timing of test administration can influence results. Tests that are administered far in advance of the work to be predicted should have evidence 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 under predict 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.

6. Confirm Tests Are Working as Intended

Once initial standards for entry into physically demanding occupations are established, they will need to be the subject of 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 mission can significantly alter the requirements of the job. And new research findings may arise 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 that is not new to the military services.

Final Thoughts

The methods for establishing physical standards for specific occupations involve the six-stage process described. 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. Each stage is 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 as well as from 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 if the testing is done improperly or as occupational tasks and equipment change over time.

Acknowledgments

We would like to thank Lt Col Mark Horner and Rennie Vasquez, our project officers in the Office of Officer and Enlisted Personnel Management within the Office of the Secretary of Defense. We are also indebted to Lernes Hebert and Juliet Beyer, who directed this office during the time this work was conducted, for their guidance and support.

We are also deeply indebted to several RAND colleagues who have contributed in various ways to this report. Barbara Bicksler assisted in the writing of the report by providing guidance in how best to convey the best-practice approaches and revising the report throughout to making it clearer and more readable. Lisa Bernard edited the report. CDR Laura Collins and Lt Col Charles Underhill, RAND military fellows from the Coast Guard and the Air Force, respectively, provided valuable feedback on topics covered in the report and continue to contribute their considerable knowledge and expertise on military occupations to the project as a whole. Kevin W. Chlebik, a student at the Pardee RAND Graduate School, Phoenix Voorhies, a RAND research assistant, and Jason Way, a RAND summer intern, assisted in gathering a variety of relevant research articles and provided their insights during internal team meetings leading up to this report. Finally, Kristie Gore, Associate Program Director for Military Health in the RAND Forces and Resources Policy Center, and Dr. Curt Gilroy provided very helpful comments on an earlier draft.

Abbreviations

CODAP	Comprehensive Occupational Data Analysis Program
DoD	U.S. Department of Defense
DOL	U.S. Department of Labor
OASD(FMP)	Office of the Assistant Secretary of Defense for Force Management Policy (since 1994, incorporated into the Office of the Under Secretary of Defense for Personnel and Readiness)
FY	fiscal year
GAO	U.S. General Accounting Office (since 2004, the U.S. Government Accountability Office)
NDAA	National Defense Authorization Act
OUSD(P&R)	Office of the Under Secretary of Defense for Personnel and Readiness
SecDef	Secretary of Defense
SEPTA	Southeastern Pennsylvania Transportation Authority
SME	subject-matter expert
VO_2 max	maximum volume of oxygen used during incremental exercise

Chapter One. Introduction

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 (SecDef, 1994) and the intention to "integrate women into occupational fields to the maximum extent possible" (U.S. Department of Defense [DoD], 2013). The rule restricted assignments of women 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 including physically demanding tasks the "vast majority" of women cannot do (SecDef, 1994).

During the next two and a half years, previously closed occupations will be opened to women who can meet occupation-specific, gender-neutral standards of performance. The SecDef has established guiding principles for implementing this policy change; these include *validating* current performance standards for military occupations, with special attention to those occupations closed to women, and *establishing* new standards where no appropriate ones currently exist.

Military service is physically demanding, and the occupations closed to women include some that are highly physically demanding. When these physical standards are in place, they will be used to match the measured capabilities of service members to the capabilities determined to be required for military occupations. Accordingly, the Secretary directed that the physical standards set for all military occupations be gender neutral. Gender-neutral standards are based only on the physical capabilities required to perform the job, are the same for men and women, and should not differentially screen out (i.e., fail to select) a higher proportion of members of one gender who are, in fact, able to perform the job. Gender-neutral standards are distinctly different from gender normed standards, in which standards are set in such a way to result in a more proportional representation by gender.

The military services have traditionally set two types of physical standards. General fitness standards have been established over the years to promote overall health status and physical fitness among military personnel.[1] These standards are not intended to ensure performance in a particular occupation. For the most part, these standards apply to all officer or enlisted personnel within a service, regardless of occupation. They need not be gender neutral. The services also set occupation-specific standards to ensure that service members are capable of performing the particular jobs to which they have been assigned. And it is these occupation-specific standards that are the focus of our study and must be gender neutral.

[1] For more on establishing military fitness standards, see Gebhardt, 2000.

The challenge in establishing occupation-specific physical standards is to determine: (1) what physical capacities are required to perform the job, (2) the most suitable tests for assessing the relevant capacities, and (3) the right minimum passing score for those tests. For the physical standards to be valid, the test assessments must measure physical capabilities required for the job and be appropriately correlated with job performance. In addition, the minimum passing score on the tests must be set appropriately. If the passing score is set too low, the standards will allow individuals into the occupation who are not qualified to perform the job and they may also be at increased risk of injury. In contrast, setting the passing score too high will result in the exclusion of individuals who are capable of performing the job and reducing the pool of individuals available to serve in an occupation. In so doing, it can also decrease the opportunity to screen on other dimensions that may be important in job performance and unnecessarily deny opportunity to individuals interested in the occupation. Therefore, an appropriate evidence base is needed to establish physical standards that screen out individuals who cannot do the job, without also screening out significant numbers of individuals who can do the job.

When selecting among tests available to assess a specific capability and both are similarly correlated with job performance, other requirements become relevant. The optimal assessment will best distinguish between those who can and cannot perform the job. Typically, there are multiple physical capabilities that must be assessed. Therefore, the merits of both the individual assessments and their collective effectiveness are relevant. In addition, the assessments must be feasible, reliable and consistent in the settings in which they need to be carried out.

Civilian employers whose jobs are physically demanding have long faced scrutiny regarding the appropriateness and equity of their standards. As DoD embarks on the process of developing gender-neutral physical standards, it can expect similar scrutiny—and, for this reason, wishes to adopt established best practices in this endeavor. To assist the military services in developing general and occupation-specific standards that are both relevant to performance and unbiased, this interim report describes the methods related to physical standard development and reviews the application of these standards to physically demanding occupations.

Why Standards?

What do we mean when we use the term *standards*? Throughout this report, we use *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.[2] Some standards are valid for screening people for a particular job, and some are not. Standards that are *valid* are those that distinguish between those who are likely to be able to perform to the requirements of the job from those who are not. Standards that do a better job at making that distinction are more valid than those that do

[2] Standards may also refer to the performance requirements of the job, or *performance standards*. In this report, if used without specific to performance, we use the word standards to refer to selection criteria for an occupation.

not. Defensible standards are those that have been developed according to best practice, that have been shown to be valid for all relevant subgroups,[3] and for which documentation describing the results of that development and validation process exists. Standards can be applied to the job itself, whereby minimum levels of acceptable performance on the job are delineated, or they can be applied to a *test*[4] intended to predict future performance on the job. In this study, we are most concerned with standards that are used to make *selection* decisions—that is, decisions made that may exclude people from entering or continuing in a job.

Much of the literature describing best practice in setting standards for physically demanding jobs draws on the experience of civilian employers, including police and fire departments. These employers are required under Title VII of the Civil Rights Acts of 1964 and 1991 (Pub. L. 88-352; Pub. L. 102-166) to develop standards that are free of bias against *protected groups*.[5] That is, if the organization chooses a method of selecting employees that results in selection of different proportions of each protected group of applicants, it must show that the standard it uses predicts the minimum level of performance required on the job regardless of group membership (i.e., that it is unbiased).

Title VII (and hence the legal protections against race and gender bias) does not apply to the selection of military personnel;[6] nevertheless, the examination of adverse impact is a crucial step in evaluating selection practices whether governed by Title VII or not. Experts in the field of personnel testing and assessment advocate the examination of test fairness for any groups that an organization wants to protect.[7] And the methods for examining impact (one element of test fairness) would be the same in a civilian or military employment context.

In addition to examining bias, personnel selection experts also recommend examining *validity* of the standards, or how well they distinguish between those who will and those who will not be able to meet the minimum requirements of the job. Validity of occupational standards is even more important for the military than it is for most civilian-sector employers because the military is expected to both spend the taxpayer's resources wisely and protect the nation from harm. Neither aim is well served if the military assigns people to jobs when their success in those jobs is highly unlikely. In the case of physically demanding jobs, many potential costs are incurred from the mismatch between skills and assignments. The following are some examples:

---

[3] Subgroups to be examined could include gender, race, or any other groups that the services would or should be concerned about excluding unfairly.

[4] In this report, *test* broadly refers to anything that might be used to exclude or disqualify someone from a job. We also use the terms *measure*, *tool*, or *assessment* interchangeably with *test* throughout the report.

[5] Title VII defines a *protected group* as individuals characterized by gender, race, color, religion, sex, or national origin.

[6] DoD is subject to Title VII with respect to its civilian employment practices, however.

[7] See the *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology, 2003) and the *Handbook of Employee Selection* (Farr and Tippins, 2010).

- *Economic costs from reduced performance on the job.* If personnel arrive on the job and cannot perform the work, then a second person may have to pick up the slack. Or a task may take twice as long or twice as many people as necessary to complete. If it takes two people to do the work expected of one, the cost to the taxpayer is doubled.
- *Lives lost because of inadequate performance.* Each person's work affects the mission. Some jobs have a much larger potential impact on people's lives. In those jobs, if something is not completed in time or to minimal standards, the loss could be grave. For example, if someone sent to rescue a downed pilot is not strong enough to carry the wounded pilot to safety, both the rescuer's and the pilot's lives could be lost.
- *Medical and disability costs due to injuries to self or others.* Even if lives are not lost because of inadequate performance, individuals may become injured or may conduct a task in such a way that other service members become injured as well. Lost time on the job that may result from injuries can affect unit readiness or mission performance.
- *Training costs lost due to attrition.* If service members are not properly screened for physically demanding jobs, the attrition rate during training may be higher than expected. If someone drops out of training, the costs spent to that point in time are wasted. And the services will have to spend additional resources identifying and training replacements, which takes time and can affect unit readiness.

All these potential costs are justification for establishing standards for physically demanding jobs regardless of who will fill the job—men or women. The recent opening of many positions to entry for women adds just one more element to consider. The challenge for the military services is to identify a set of standards to address each of the above costs that are the same regardless of gender—that is, standards that are *gender neutral*.[8] The standards also should be consistent with the goal of expanding opportunity for women to enter occupations for which they are qualified, as stated by the Secretary of Defense (Panetta, 2013) in eliminating the direct ground combat exclusion rule:

> The chairman and the Joint Chiefs of Staff and I believe that we must open up service opportunities for women as fully as possible. And therefore today, General Dempsey and I are pleased to announce that we are eliminating the direct ground combat exclusion rule for women and we are moving forward with a plan to eliminate all unnecessary gender-based barriers to service.
>
> 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.

This report reviews how to establish standards based on the requirements of physically demanding jobs using best-practice methods. It is also useful to note that physical standards can

[8] Although we acknowledge that the social sciences often distinguish between the terms sex and gender, in this report the terms are used interchangeably to refer to the same concept.

be and are commonly applied at multiple career points. Standards determine who qualifies to enter training for the occupation, but the training curricula also set standards that qualify individuals to graduate from training and enter the occupation,[9] and standards may be established for determining who will be allowed to continue in the occupation later in their career.s The issues of gender neutrality and the best-practice methods described in this report apply equally to entry, training, and on-the-job standards.

This report is about the proper methods for establishing physical requirements for jobs. In addition, because of the recent DoD decision to open ground combat roles to women, we also discuss some gender-related policy issues associated with establishing those physical requirements. But ensuring that standards are gender neutral is only one part of a much larger process. The basic principles, methods, and policies discussed throughout are intended to apply equally to jobs that have been already open to women for decades, were just recently opened, or may be opened in the near future.

Occupation-Specific Physical Standards and Entry into Military Service

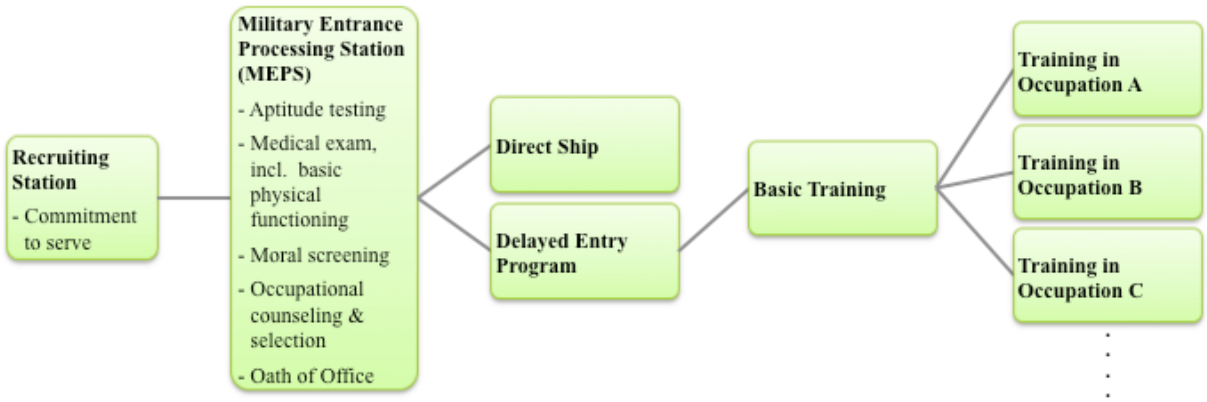
To develop appropriate tests and cutoff scores for screening and selecting new entrants for physically demanding military occupations, it is important to determine at which point in the entry process the screening will be done and standards for entry into the occupations determined. Figure 1.1 depicts the major steps all enlisted personnel take as they enter service, regardless of service or occupation. The figure focuses on enlisted personnel who enter occupational training immediately after completing basic training. In contrast, most officers enter their occupations only after completion of several years of preparatory training, at the military academies or in Reserve Officer Training Corps programs at colleges and universities. In theory, screening of entering enlisted personnel can be done at several points: before the individual commits to enlisting at the recruiting station, before he or she is placed in an occupation at the Military Entrance Processing Station (MEPS), or at arrival or completion of basic training.

Deciding when to conduct occupation-specific physical screening involves some clear trade-offs. On the one hand, individuals improve their physical capabilities in basic training, but the level of improvement varies and is difficult to predict. Screening for eligibility to enter physically demanding military occupations is likely to be more accurate if done at the end of basic training than screening at any of the earlier stages. Screening at the point in time when the tests are most predictive of occupational training success should decrease attrition from the resource-intensive occupational training programs. Waiting until the end of basic training to screen for entry into occupational training would catch most individuals who are not capable of

[9] To the extent that attrition occurs because of someone's inability to meet the physical requirements in training or on the job, standards are being applied even if they have not been formally established as career-specific physical standards.

meeting the physical demands of the occupation, but there would be limited time to direct individuals who fail to qualify to other occupations. Therefore, late screening could increase the number of enlistees who must be sent home.

Figure 1.1. Enlistment and Initial Training of Military Enlisted Personnel



On the other hand, both the individual enlistee and the services benefit from screening early in the process, before they enter the training pipeline. The individual learns something about the physical demands of different occupations and his or her physical capabilities before completing the enlistment process and likely faces a smaller risk of failing to meet the requirements to enter occupational training. The services benefit by more accurately managing the flow of recruits to basic training, taking into account the availability of occupational training seats.

When the screening will be implemented must be determined before the tests and standards are developed. As we discuss in the remainder of this report the occupation-specific screening tests and eligibility standards (e.g., minimum scores required to enter the occupation) are determined based on analysis of the relationship between performance on the screening test and performance in occupational training or subsequently on the job. Key elements of the analysis must be carried out with test subjects who accurately represent the population of enlistees at the point in the process at which the screening will be done. Otherwise, the wrong tests may be selected, and the eligibility standards may be set at the wrong level.

Determining the optimal point in the enlistment process for occupation-specific screening would take considerable time and analytic resources. However, once initial occupation-specific screening and standards are implemented, the services can explore over time whether implementing them elsewhere in the enlistment process would add benefit.

Study Approach

The Office of the Under Secretary of Defense for Personnel and Readiness (OUSD(P&R)) asked RAND to help it understand how to evaluate job-specific physical requirements and establish gender-neutral standards for physically demanding jobs. Our study addresses two research objectives. First, in this report we describe the methodologies for establishing the standards for physically demanding jobs, tailored to address the needs of the military.

The second objective of this study is to review and evaluate methodologies being used by the military services to set gender-neutral standards. The results relating to this second objective are presented in a separate report. That report uses the concepts presented here as a framework for reviewing the services efforts to establish and validate their standards.

Organization of This Report

The remainder of this report provides the results of work toward the first research objective:

- Chapters Two through Eight review the methods for establishing and validating evidence-based standards. These chapters discuss methods for identifying a job's physically demanding tasks; selecting an appropriate set of screening tests for further consideration; determining which tests are most useful for predicting important organizational outcomes regardless of gender; setting minimum scores on the tests; and establishing an ongoing data-collection and analysis process to ensure that physical requirements are current and have been accurately assessed.
- The final chapter summarizes the key steps in the standards development process and describes next steps in conducting our study.

Chapter Two. Methodological Approaches to Establishing Physical Job Requirements

Methods for establishing requirements for physically demanding jobs combine insights from two main disciplines: personnel selection and physiology.

The professional practice guidelines in the field of *personnel selection* are well established as the primary source regarding the proper use and development of tests and measures in employment contexts. They are also the basis for much of the content discussed in the federal government's Uniform Guidelines on Employee Selection Procedures (Code of Federal Regulations, 1978).[10] The Uniform Guidelines do not apply to the military, but they may serve as a reference in ensuring that the goal of "eliminat[ing] all unnecessary gender-based barriers to service" is met. In addition, an overview of professional practice guidelines (established independent of Title VII and the Uniform Guidelines but also used to inform them) for developing and evaluating employment selection measures can be found in two published resources:

- *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology, 2003). This source (referred to as the *Principles*) was produced by the Society for Industrial and Organizational Psychology to "specify established scientific findings and generally accepted professional practice in the field of personnel selection psychology in the choice, development, evaluation, and use of personnel selection procedures designed to measure constructs related to work behavior with a focus on the accuracy of the inferences that underlie employment decisions" (p. 1).
- *Standards for Educational and Psychological Testing* (Joint Committee on Standards for Educational and Psychological Testing, 1999). This source (referred to as the *Standards*) was developed jointly by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. It summarizes professional standards for the development and use of tests in educational, psychological, and employment settings. According to the Department of Labor (DOL), the standards "are consistent with applicable regulations and are frequently cited in litigation involving testing practices" (DOL, 1999).

Although many of the guidelines in the *Standards* are directed at assessing mental knowledge, skills, and abilities, the same measurement concepts apply to the assessment of

[10] The *Uniform Guidelines*—adopted by the Equal Employment Opportunity Commission, the U.S. Department of Labor (DOL), the U.S. Department of Justice, and the U.S. Civil Service Commission—are "intended to establish a uniform Federal position in the area of prohibiting discrimination in employment practices on grounds of race, color, religion, sex, or national origin" (29 CFR Part 1607; 41 CFR Part 60-3; 28 CFR § 50.14, 5 CFR § 300.103[c]).

physical skills and abilities. This applicability is noted explicitly in the *Principles*. The methodological approaches that we describe are consistent with those advocated in both the *Principles* and the *Standards*.[11]

The second domain playing a central role in establishing requirements for physically demanding jobs is physiology, which offers a vast literature on anatomy, injury, measures of physiological functioning, physiological sex differences, and other domains relevant in addressing key workplace issues. For example, the field has valuable insights into how jobs can be reengineered to reduce injuries, how to reduce training injuries, and how to measure physical fitness. Most important, it can provide insights into the types of tests that might be useful for employment screening and selection.

Whereas personnel selection offers the methodological approach, physiology serves as the starting point for much of the content applied in the methodology.

Six-Stage Process

We organize the overall approach for developing physical standards in six general stages, as depicted in Figure 2.1 and described below. These steps also provide a useful framework for evaluating any standards that are already in place. For those standards the services should review their existing evidence in support of each of these steps and consider supplementing their past efforts if any gaps in the evidence are identified.

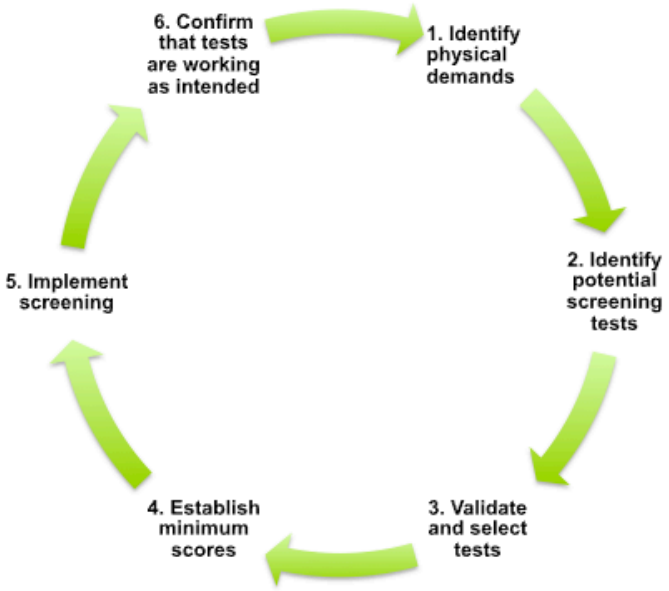
- **Stage 1. Identify the physical demands of the job.** Define all tasks required on the job, and identify which of those tasks are physically demanding and which are not. Identify other relevant aspects of performance, such as injuries, that may be affected by physical ability.
- **Stage 2. Identify potential screening tests.** Explore past research on potential screening tests, articulate reasoned theories regarding the applicability of a particular tool, and identify varied options for inclusion in validation. If standards already exist, stage 2 might appear not to be necessary. However, to guard against the possibility that standards based on these tests prove not to be valid, we recommend including other potential tests as well as the existing ones.
- **Stage 3. Validate the tests, and select those with highest validities and least adverse impact.** Administer a range of tests to job candidates, and examine the relationship between test scores and important outcomes on the job (e.g., job performance, injury rates, productivity). From the results of validation studies, identify the best predictors of performance. This step also involves analysis of adverse impact on selection within relevant population subgroups to confirm that the tests are equally

[11] The suggestions provided here and in the following chapters are generally consistent with the recommended approaches for defining requirements for physically demanding jobs in other reputable sources. See, for example, Campion, 1983; Sharkey and Davis, 2008; Hogan et al., 1979; Gebhardt and Baker, 2010a, 2010b; Baker and Gebhardt, 2012, and Arvey et al., 1992.

valid for all groups. As we discuss below, this does not mean that the pass rates should be the same for all groups, but that the tests should predict performance on the job equally well for all groups.

- **Stage 4. Establish minimum scores.** Apply a systematic process to identify minimum test scores that should be established for entry into or continuation in a job.
- **Stage 5. Implement screening.** Establish a systematic method of test administration. Train personnel in applying that method, and begin screening personnel using the test.
- **Stage 6. Confirm that the tests are working as intended.** Verify whether test administration in practice adheres to the guidelines that were established. Determine whether job requirements have changed. Examine whether coaching or test-preparation activities have compromised the test's validity. Reexamine predictive validity and adverse impact of the test.

Figure 2.1. Six Stages in Developing Physical Standards



Each element in the process for establishing physical job requirements provides support for the use or exclusion of a set of selection procedures. The deliberate steps in this process and, importantly, the documentation of the actions taken are critical in developing defensible physical standards. Among employment tests, physical tests have generated the highest number of civilian court cases and have one of the lowest rates of successful defense (Terpstra, Mohamed, and Kethley, 1999) in large measure because best practices have not been followed. In the following chapters, we provide an overview of well-accepted approaches in addressing each stage of the process. We conclude each chapter with a table summarizing key considerations we drew from the literature and the potential approaches for addressing each consideration.

Chapter Three. Identify the Physical Demands of the Job

The process for establishing an accurate accounting of the tasks or activities that take place in a job is known as *job analysis*.[12] The results of job analyses serve as the foundation for nearly all human resource management activities. They can be used to write job descriptions, design training content, classify jobs into job families, merge two jobs that have similar tasks, redesign a job, define performance expectations, adjust compensation, create performance evaluation tests, and more. Although job analysis has applications in many other contexts, it is also used to support decisions about an organization's selection system.

Job analyses already exist for many military jobs. However, unless they were developed with a focus on assessing the physical requirements of the jobs, they may contain less information than needed (e.g., level of effort required, weight of certain key objects, duration of the activity). Before moving to the next stages of the standard setting process, the job analyses should be carefully reviewed and revisited if necessary.

Methods for Conducting a Job Analysis

There are a variety of methods for collecting job analysis information, and each method produces different data.[13] The following are among the best-known methods:

- Task inventories, such as the Comprehensive Occupational Data Analysis Program (CODAP) system (Christal, 1974), produce a detailed and comprehensive list of tasks performed on the job and ask a representative sample of job incumbents to rate the task on such factors as importance and frequency with which the tasks are performed.
- The critical-incident technique (Flanagan, 1954) also can be used to generate detailed task statements by asking subject-matter experts (SMEs), who are often job incumbents, to describe an incident that shows exemplary or poor performance, the events leading up to the incident, and the resolution or outcomes resulting from the incident. Those incidents can then be used to create a job incumbent questionnaire similar to that produced by CODAP.
- Functional job analysis (Fine and Getkate, 1995) focuses less on documenting a comprehensive list of tasks performed on the job and more on documenting what workers do in relation to three key elements on the job: people, data, and things. This is the

[12] Other terms used to refer to the same systematic processes of defining jobs include *task analysis*, *occupational analysis*, and *work analysis*.

[13] See Gael (1988) and Brannick, Levine, and Morgeson (2007) for additional information on how to conduct a job analysis.

approach used to develop DOL's Occupational Information Network (O*NET) database of occupational requirements and worker attributes.[14]

- The Position Analysis Questionnaire (McCormick, Jeanneret, and Mecham, 1972) relies on a predetermined set of questions that are the same regardless of the occupation. The questions cover a wide variety of topics, including the environment in which the work is performed, the types of information sources used on the job, mental processes, and work output. The result may be highly detailed, but it does not provide task descriptions that are unique to an occupation.

Job analyses can focus on collecting distinctly different types of information. Some are worker-oriented approaches, which focus on "what workers do in performing their jobs (e.g., visual, manual, or communication activities), while others are job-oriented approaches, which focus on "what workers accomplish in their jobs" (e.g., baking, selling, painting) (Palmer and McCormick, 1961). The Position Analysis Questionnaire is one example of a worker-oriented approach. Functional job analysis, task inventories, and the critical-incident technique are examples of job-oriented approaches.

The mode of data collection can also vary widely. Data could be collected through observations of people performing the job, through focus groups or interviews, or via paper-and-pencil or online questionnaire. The people who serve as experts range as well. Sometimes, job analysts serve as the SMEs; in other cases, they call on job incumbents, supervisors, scientists, or training instructors to provide expertise. In many studies, more than one method of data collection is used and more than one type of expert is consulted. For example, in methods involving an occupation-specific questionnaire, focus groups with job incumbents may be used to develop the tasks on the questionnaire. In jobs that are new, supervisors or instructors for the job might be consulted to identify the tasks or challenges incumbents are likely to face in the future.

There is no single correct choice among these methods of job analysis (Gael, 1988; Brannick, Levine, and Morgeson, 2007). Any of them may be appropriate and adequate in some circumstances. Choice of one method over another and decisions about the level of detail that is necessary should instead be driven by the goals for the use of the results.[15] However, it is important to note that job analysis data that are adequate for one activity may not be adequate for another. For example, a job analysis designed solely for creating a short job description could produce far less-detailed information about the job than a job analysis designed to define the content of a comprehensive job training program.

In establishing gender-neutral requirements for entry into physically demanding jobs, the focus is on applicant selection. In this study, the intended use of the job analysis is to design a

---

[14] Available online at **http://online.onetcenter.org.**

[15] When the results of a job analysis are intended for use in multiple personnel activities, all of those uses should be considered in determining the appropriate methodology or methodologies.

selection system for physically demanding jobs. The primary goal of the job analysis, therefore, should be to identify and describe in detail the physically demanding tasks the applicants would need to be able to perform in the job. Task-level detail specific to each occupation is ideal for a sound defense of a selection system.[16] The more information and detail on the job's physical demands, the better.

Although there is no single appropriate methodology, attending to several key features is important. One is the choice of SMEs. Experience level of the SMEs can have a meaningful impact on the results because less experienced personnel are typically less knowledgeable about the particular contents of the job. Relying on supervisors instead of incumbents may fail to capture important features of the job as it is actually done. Other important factors are the number and seniority of people involved in the job analysis. The involvement of only a few people from only a few locations may not sufficiently represent the overall worker population and therefore may mask important variation in the job. Similarly, the involvement of only senior personnel could fail to capture important work duties performed only by junior personnel. In the case of jobs in which one group (for example, women) is underrepresented, it may be relevant to ensure that sufficient numbers from that group are included to allow for comparison of the results by group.

Reasons to Conduct a Careful Job Analysis

The centrality of a job analysis in defending the use of a selection system cannot be overstated. Without an accurate understanding of the content of the job, a sound argument supporting a given selection tool cannot be made. The job analysis is fundamental to ensuring that the standards for an occupation are valid predictors of critical job requirements. Although Title VII does not apply to the military, it can provide insights into the importance of job analysis in developing appropriate standards, and the courts clearly view it as important. According to Landy and Vasey (1991),

> In virtually all Title VII cases litigated at the Federal level, there is an extensive examination and discussion of the job analysis techniques that were or should have been used in the particular validity study. . . . Most commonly, plaintiffs will assert that there is a fatal flaw in the job analysis techniques, analyses, results, or inferences. They may assert, for example, that important or frequently performed duties were ignored or that unimportant or infrequently performed duties were given too prominent a role in test development. For their part, the defendants will commonly rebut that charge by suggesting that there is no one acceptable method of conducting a job analysis and that the analyses, results, and inferences are appropriate and support the identification or development of the selection strategy being considered. (p. 29)

[16] See Hogan and Quigley, 1986, for a discussion of the types of job analysis techniques that have been successfully defended in past court cases.

Thus, the choice of methodology used to define the content of the job can be vital in addressing some of the criticisms that might be raised in the context of physical testing. The job analysis should take into consideration a variety of factors (in Table 3.1, we provide concrete examples) and, if applicable, take steps to ensure that the job analysis is appropriately complete for the circumstances. Without taking these considerations into account, organizations can be open to criticism regarding the efficacy of the job analysis.

Table 3.1. Example Considerations in Identifying the Physical Demands of the Job

Considerations	Potential Resolution
Are non–physically demanding tasks (and therefore other skill sets) more important than the physically demanding ones?	Include in the job analysis all typical and all important tasks performed on the job (not just those that are physically demanding). Assess how important each task is to the job and how frequently it occurs on the job.
Do the physically demanding tasks occur infrequently, or are they not demanded of everyone?	Assess how frequently the tasks are performed on the job. Identify what proportion of workers has to perform each task.
Does the type of physical skill required on the test reflect the skills required on the job? For example, a test of upper-body strength is not the right test for a job that requires mostly trunk or lower-body strength.	Produce a detailed description of each job task, including the objects involved (e.g., ammunition can), the physical movements involved (e.g., lift to height of truck bed), and types of equipment used (e.g., hand truck).
Is the level of performance required on the test higher than that required on the job?	For physically demanding tasks, (1) The actions involved should be defined clearly (e.g., lowering a 75-pound explosive into a missile silo). (2) The weights of objects and the duration and frequency of the tasks should be determined. (3) The level of effort and speed expected for task performance should be determined. (4) The number of people and types of devices (e.g., the use of a hand truck) that typically provide assistance in performing the physically demanding tasks should be identified.
Are individuals who are selected as experts for the job analysis sufficiently knowledgeable? Is the number of experts too small or not representative of the job as a whole? Are variations in the job adequately considered—e.g., differences from location to location, differences in tasks in higher- versus lower-level positions, or alternative ways to accomplish the same task?	Include participants from a wide variety of locations, at all levels of the job, and a sample that ensures representation of the occupation as a whole. Compare the responses across locations and across job levels to determine whether the job differs by location or job level. Examine the variability in responses across all respondents to determine how the activities on the job vary from person to person.
Are the job incumbents used as SMEs for the job analysis biased? Is there a culture of competitiveness that might lead individuals to exaggerate the importance or level of the physical demands?	Have a panel of independent experts review and evaluate the appropriateness of the activities described for accomplishing the mission. Include a diverse but appropriate set of experts (e.g., include women and men who have experience working in a related field).
Should the job be modified to accommodate more people? Could it be reengineered to reduce the physical demands?	Ask participants which tasks could easily be modified to allow more people to perform them successfully (e.g., buddy system for lifting, hand trucks).
Did the job analysis examine whether women might accomplish physical requirements of the job in a different but equally effective way?	Explore ways to include a sample of women in the job analysis process and to examine gender differences in how the activities are performed and the importance of the activities. If an insufficient number of female job incumbents are available (as would be the case in jobs previously closed to women), consider bringing in a panel of qualified women (e.g. experienced in other physically demanding occupations) to observe and learn about the job and include their perspectives in the job analysis.

The results of a job analysis can also—if designed with this in mind—set the groundwork for other stages in the process of establishing requirements. For example, it could be designed to support an argument that simulation activities during training are good approximations of how well people will perform important tasks on the job. If such an argument can be made successfully from the contents of the job analysis, then performance in the training simulations could be used as an outcome measure in a predictive validation study (see Chapter Six for more on this). Although a job analysis that addresses the issues listed in Table 3.1 would likely be useful for designing training simulations as well, we caution users to think critically about what can and cannot be extrapolated from each job analysis, particularly when using the results of a job analysis for a purpose other than the one that was originally intended.

Chapter Four. Identify Potential Screening Tests

Identifying potential screening tests to measure the physical skills needed to perform job tasks is the next step in developing physical standards. As we indicated in Chapter Two, even where there do exist physical selection standards, it may be valuable to include potential tests other than the ones in use to ensure validity of the standards. Many factors weigh into identifying potential tests, but one important consideration is whether research and theoretical support exist for a tool's use in a similar employment context. The universe of tests is potentially infinite and, although research has tapped only a subset of that universe, there is a body of literature summarizing research on a variety of existing measures. Test developers and employers should be aware of the results of that research, especially when alternative tests have been shown to differ in their validity across occupations or work environments and show adverse impact against key population subgroups.[17] In cases in which an employer chooses to devise a new test, one for which research does not already exist, a clear rationale for believing it to be better than existing tests is needed and should be documented. Regardless of whether new tests are being explored or well-established tests are being considered, test developers should be cognizant of the prevailing theories involved with the measurement of physical skills.

Cost, feasibility, and applicant reactions are also reasonable considerations in selecting measures. All of these considerations are discussed further in this chapter.

Taxonomies of Physical Aptitudes

One of the most critical theoretical issues faced by researchers studying physical attributes for employee selection is how to separate the various types of activities required on the job. There is no single taxonomy of physical abilities that best addresses this issue. However, the most commonly cited taxonomy is the one devised by psychologist Edwin Fleishman in the 1960s (Fleishman, 1964). Fleishman's taxonomy was initially defined using two samples of Army recruits, and was further refined in a series of subsequent studies surveying the job performance domain. Continued work examining the underlying structure of the physical domains provides additional support for its use.[18] The concept underlying Fleishman's research is that people can score high on one physical aptitude without necessarily scoring highly on

[17] Jobs differ in type, level and importance of the physical abilities required. Consequently, the most appropriate physical-ability test will depend on the job. Research on jobs that are similar can and should be used to inform which tests would likely be the most appropriate and have the least adverse impact; however, validation research must still be undertaken to confirm applicability and usefulness for each organization.

[18] For a review, see Myers, Gebhardt, Crump, and Fleishman, 1993.

others. Selecting the right tests in an employment context, therefore, requires careful attention to which physical abilities are and are not required by the job.

Fleishman divides physical abilities into five general areas: strength, flexibility, coordination (or agility), equilibrium (or balance), and stamina (cardiovascular). And each area defines basic abilities or domains (Table 4.1). Some research has shown that, at the most basic level, physical skills can be grouped into fewer factors than are outlined in Fleishman's taxonomy. For example, Hogan (1991) showed that the physical demands of the job can be summarized with just two broad factors, while physical ability tests can be grouped into three broad factors.

Table 4.1. Fleishman's Physical Ability Domains

General Area	Basic Ability or Domain	Domain Definition
Strength	Dynamic strength	Ability of the muscles to exert force repeatedly or continuously over a long time period. This is the ability to support, hold up, or move the body's own weight or objects repeatedly over time. It represents muscular endurance and emphasizes the muscles' resistance to fatigue.
	Trunk strength	Involves the degree to which one's abdominal and lower-back muscles can support part of the body repeatedly or continuously over time. The ability involves the degree to which these trunk muscles do not fatigue when they are put under such repeated or continuous strain.
	Static strength	Ability to use muscle force in order to lift, push, pull, or carry objects. It is the maximum force that one can exert for a brief period of time.
	Explosive strength	Ability to use short bursts of muscle force to propel oneself or an object. It requires gathering energy for bursts of muscle effort in a very short time.
Flexibility	Extent flexibility	Ability to bend, stretch, twist, or reach out with the body, arms, or legs as far as possible in a forward, lateral, or backward direction
	Dynamic flexibility	Ability to bend, stretch, twist, or reach out with the body, arms, or legs, both quickly and repeatedly
Coordination (agility)	Gross body coordination	Ability to coordinate the movement of the arms, legs, and torso together in activities in which the whole body is in motion
Equilibrium (balance)	Equilibrium	Ability to keep or regain one's body balance or stay upright when in an unstable position. This ability includes maintaining one's balance when changing direction while moving or standing motionlessly.
Stamina (cardiovascular)	Stamina	Ability of the lungs and circulatory systems of the body to perform efficiently over long time periods. This is the ability to exert oneself physically without getting out of breath.

SOURCE: Industrial/Organizational Solutions, 2010, pp. 4–5.

However, it may be important to consider even finer distinctions than those in Fleishman's taxonomy when choosing tests for use in a personnel selection context. For example, Myers Gebhardt, and Fleishman (1980) argued that, because lower-body versus upper-body strength differ by gender, it would be important to measure them separately when conducting research to support a selection measure. Using a sample of four Army occupations, they demonstrated that

job analysis questions to evaluate upper-body and lower-body strength separately within each of the four strength factors were reliable and did distinguish between the two aspects of strength. There is also evidence showing that the applicability of a test for measuring a given area of the taxonomy can, in some cases, differ by gender. For example, Myers, Gebhardt, Crump, and Fleishman (1993) found that level of body fat was a significant predictor of physical test scores for men but not for women. In selecting standards that apply equally to men and women, the military should be cognizant of potential differences by gender such as these.

Examples of Tests Studied for Use in Employment Settings

Researchers have considered a wide variety of tests for use in employment settings. Some have been empirically investigated in the research literature for use in employee selection. Table 4.2 provides examples of these tests, by domain.

Table 4.2. Examples of Tests Used to Measure Physical Abilities in the Different Domains

General Area	Basic Ability or Domain	Test
Strength	Dynamic strength	Push-ups Pull-ups Flexed arm hang
	Trunk strength	Leg-lifts Sit-ups Hold half sit-ups
	Static strength	Hand grip
	Explosive strength	60-second box jump Softball throw Standing broad jump
Flexibility	Extent flexibility	Sit and reach Shoulder reach flexibility test
	Dynamic flexibility	Lateral bend One-foot tapping test
Coordination (agility)	Gross body coordination	Illinois agility 505 agility
Equilibrium (balance)	Equilibrium	Stork stand
Stamina (cardiovascular)	Stamina	Multistage fitness Step test

Selecting Candidate Tests

A variety of factors come into play when selecting candidate tests to measure the physical abilities necessary to perform a particular job. Three of the most important are fidelity to the job, cost, and feasibility.

Fidelity to the job refers to the similarity between the test and job tasks. High-fidelity tests have obvious overlap with the job. Examples include simulations or work samples, such as asking firefighter candidates to perform a variety of typical firefighter tasks, such as carrying a hose for a specified distance or carrying a dummy down a ladder, or asking commercial pilot candidates to take off and land a plane in a flight simulator. These tests often can predict job performance. Low-fidelity tests, in contrast, are those that have little observable similarity to the job tasks. Instead, they measure more-general physical abilities that may be relied on to perform job tasks. For example, measures of oxygen uptake (such as VO_2 max, the maximum volume of oxygen used during incremental exercise) or hand-grip strength are highly abstract relative to the tasks of most physically demanding jobs (e.g., firefighting, rescue swimmers), although they may still be valid predictors of success in physically demanding tasks for those professions. There can be some overlap in the two types of tests, and either type or a combination of both types can be effectively used to screen job applicants. The choice may vary across occupations.

High-fidelity tests offer some benefits over low-fidelity simulations. For example, tests that have obvious overlap with the job are viewed as more *face valid*[19] and therefore fairer by test takers, reducing the likelihood that applicants will challenge the test. If the test does face legal challenge, a well-documented job analysis that supports fidelity to important or frequent job tasks should be sufficient to defend its use. (This is discussed further in Chapter Six's discussion of content validity.) However, some high-fidelity tests can be costly to develop and administer, and validity arguments based solely on content overlap with the job may not support the tests' use for occupations that do not share the same job tasks.

Cost is an important factor when selecting tests. This includes equipment costs (e.g., cost of purchasing, operating, and replacing equipment; facilities to house the equipment or the testing location), manpower costs (e.g., applicant time, test administrator time, time to train test administrators, costs of scoring the tests), validation costs (cost of conducting research to support the test's use), and perceived fairness costs (which could range from minor psychological costs, e.g., reduced organizational commitment, to major resource expenditures, e.g., litigation).

Tests can vary widely in potential cost. For example, a treadmill, although easily accessible at most fitness centers, is expensive to purchase if the test will be conducted where facilities do not already exist, and the time required to administer and complete some types of treadmill tests (e.g., time it takes to reach the point of exhaustion) is not insignificant. Less expensive alternatives that may produce essentially the same information should be explored.

---

[19] Face validity is determined solely by lay perceptions of the test and may be entirely unrelated to the actual validity of a test. For example, a valid predictor of later performance might not appear face valid to test takers, or a test might appear face valid even if it is not a valid predictor at all. Because face validity is unrelated to actual validity, it does not qualify as evidence-based support to justify a test's use. Face validity does matter, however, when considering test taker's perceptions of test fairness.

Feasibility is a third consideration, one that is closely aligned with cost. For example, given limited time and resources, some tests cannot easily be administered across multiple locations with accuracy or consistency. To illustrate, for a job with locations around the country, a timed swim test of all applicants would face some logistical challenges. It would first require that all locations have access to a pool for testing or that all locations send their applicants to a central location for testing. If the test is administered locally, administration protocols and scores would need to be adjusted for each pool's distance per lap because a lap from one pool may not be comparable to a lap in another pool.

Cost and feasibility would be of particular concern if the services wanted to scale up an occupation-specific simulation for use by recruiters. Most tests are likely to be too complicated to replicate with accuracy at multiple locations without dedicating significant resources. Even if a simulation were replicated at multiple locations, the prospect of doing so for multiple military occupations is not likely to be practical. Acquiring the facility space to accomplish such broad testing would be daunting, not to mention the costs of paying people to observe and score applicants.

The military faces a unique challenge in selecting a set of tests for initial job classifications. There are many different military occupations for which a screening test would be useful. But administering a high-fidelity simulation to all military applicants would be time and cost prohibitive. Instead, administering a series of simple tests that can generalize across more than one job would be a more feasible approach. However, simulations can still be a feasible approach to screening. Simulation activities could take place during basic training or occupation-specific training to eliminate people from the career field. This is more feasible because the simulation would need to be only in the limited number of locations where training already occurs.

In Table 4.3 we provide examples of factors that should be considered in selecting types of tests and ways to resolve them.

Table 4.3. Example Considerations in Identifying Potential Screening Tests

Consideration	Potential Resolution
Do the tests cover all of the physical ability dimensions relevant for the job? Do the tests tap physical ability dimensions that are not relevant for the job?	Establish and document a solid rationale based on the relevant research literature for selecting various tests. Use the information gained from job analysis and existing research to support that rationale. Ensure that the rationale does not conflict with past research or theory supported in the research literature.
Do other, more valid tests exist?	Include a variety of tests in the validation process, and document the process for selecting them. Conduct a comprehensive review of the types of tests that could be used, and identify those with greatest promise based on existing research and theoretical grounds. Consider other tests that may not have been well studied but have theoretical merit. Document all tests that were considered and why each test was included or excluded from the final set to be validated. In the decisions, consider existing research on the tests' validity, fidelity to the job, cost, and impact on selection by population subgroup. If cost is used as a reason to exclude one expensive test but not another, document and explain the inconsistency.
Are the tests feasible and cost-effective?	Document which tests are considered not feasible or cost-effective and how that determination was made. Weigh feasibility and cost-effectiveness in the decisions about which tests to include in the validation studies.
Do other tests have lower adverse impact on women?	Include tests that could be useful predictors but are known to have smaller race or gender differences. Compare validation results with those that have larger race or gender differences.

Chapter Five. Validate and Select Tests

The third step in developing physical standards is to validate potential tests and identify those with the highest validity and least adverse impact. The word *validate* is often used loosely to refer to any process intended to check or confirm the correctness of a policy or practice. Accordingly, the act of asking an organization's leadership or an expert to sign off on a screening tool's use is often referred to as *validating* the tool. This is not, however, consistent with the meaning of *validation* as it is defined in the context of personnel selection.

In the personnel selection context, the term *validate* has a much more 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 (Messick, 1980, 1989, 1995; Anastasi, 1986; Binning and Barrett, 1989; 29 CFR Part 1607; 41 CFR Part 60-3; Society for Industrial and Organizational Psychology, 2003).[20] The following are three key types of validation evidence:

- Evidence of *content validity* is evidence that a test covers the job content domain of interest.
- Evidence of *criterion-related validity* is evidence that the test predicts important organizational outcomes.
- Evidence of *convergent* or *discriminant validity* is evidence that a test measures what it purports to measure.

This chapter describes each type of evidence in the context of developing occupation-specific physical standards. Multiple sources of evidence should be accumulated to demonstrate whether a test measures what it is intended to measure and that its scores can be used for selection.[21] Each piece of content, criterion-related, and convergent or discriminant validation evidence that an organization collects lends additional support to that determination.

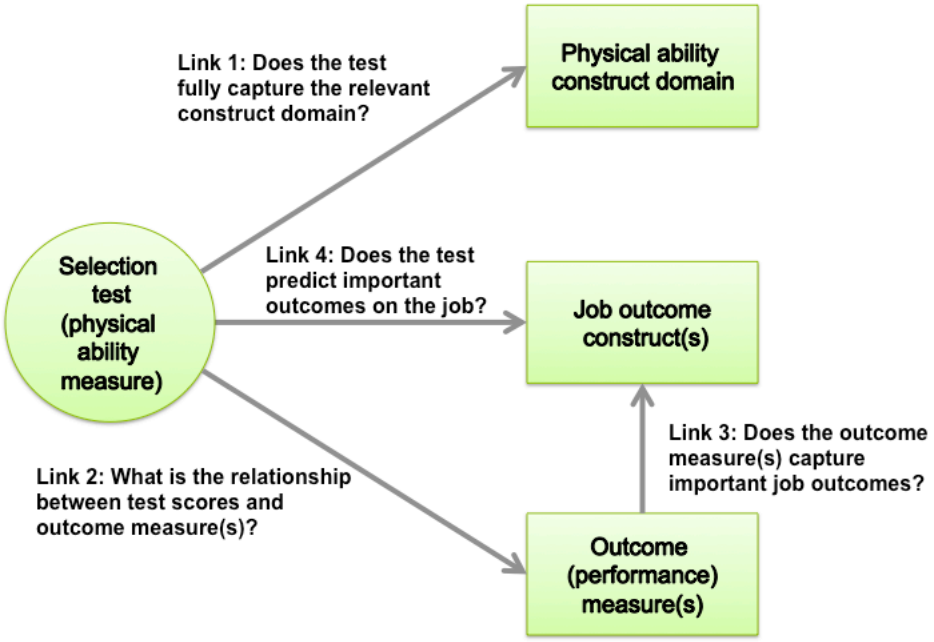
Figure 5.1 illustrates the conceptual linkages that can be examined during the validation process. In personnel selection, the ultimate goal of validation is to provide evidence to support link 4, that the selection test predicts important outcomes on the job. No single method of validation can provide complete support for that link. Instead, amassing information that

[20] The process of collecting multiple sources of research-based evidence to support a tests use is also referred to as *construct validation* (for more on the meaning of *construct validation*, see Anastasi, 1986.

[21] For example, there would be strong theoretical support for the use of a test that is backed up by two different well-designed pieces of criterion-related validity evidence (e.g., prediction of training success and performance in a realistic job simulation six months after training) and a well-designed study of content validity. Such a combination provides three pieces of evidence to support the test's use. A test for which there is only one piece of evidence (such as one estimate of criterion-related validity) would still have support; however, that support would not be as strong. The greater the variety of study designs and evidence that is amassed, the stronger the support.

confirms all four conceptual links helps add confidence that link 4 is also supported. How each type of validation evidence relates to links 1 through 4 is discussed more in this chapter.

Figure 5.1. Conceptual Validation Linkages



SOURCE: Adapted from Binning and Barrett, 1989.

Construct Deficiency and Construct Irrelevance

The first step in any validation effort is to clearly define the *constructs* (i.e., the concepts or characteristics) one intends to measure. Verbal and mathematical aptitude, personality, job performance, finger dexterity, and physical strength are examples of broad constructs that have been explored in personnel research. Validation, however, requires development of much more-precise definitions. Precise and well-documented definitions are necessary for determining whether the test selected is a good predictor of the construct being measured—that is, whether link 1 in Figure 6.1 is supported.

Construct deficiency and *construct irrelevance* are two key concepts related to whether there is good support for link 1 in Figure 5.1.[22] A test is *construct deficient* when it fails to capture an important element of the construct domain (see Figure 5.2). For example, a high-school algebra test that does not include any equations with exponents would be construct deficient. It fails to capture an important element of the domain of high-school algebra. Similarly, a test purported to

[22] See Messick, 1989, for more on these concepts.

measure strength that measures upper-body strength but not core or lower-body strength would also be construct deficient unless. The stated scope of the test matters. A test described as measuring strength should cover the entire domain of strength. A test described as measuring the domain of upper-body strength would not be expected to tap lower-body strength as well (if it did, that part of the test would be construct irrelevant; see Figure 5.3).

Figure 5.2. Construct Deficiency

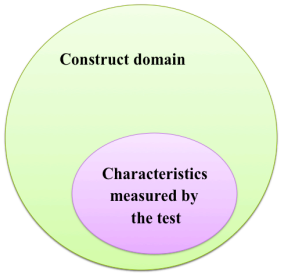
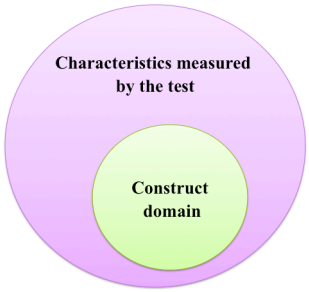


Figure 5.3. Construct Irrelevance



Use of a construct-deficient test could lead to incorrect conclusions regarding someone's competence in the domain of interest. If a construct-deficient test is used for selection, then candidates may be selected who are not capable of performing on the job (i.e., *false positives*) or candidates may be rejected who would have been capable of performing the job (*false negatives*) in higher numbers than might otherwise be the case.

Construct irrelevance (or *construct contamination*—the measurement of something other than what was intended) (see Figure 5.3)—is also problematic. It too can lead to an increased number of false positives and false negatives.

Many factors could cause construct-irrelevant variance in test scores. For example, the test administration environment might change test takers' motivations to perform well. They might perform very differently in front of a group of people cheering them on than if they had a group of silent onlookers or when no one else but the test administrator is in the room. Other types of

motivation could also affect scores. Those who want to avoid jobs that have physical demands might intentionally underperform on the test. In these examples, the resulting test scores might measure the underlying construct domain that the test was intended to measure, but they would also measure motivation. If the test were argued for use in selection on the theoretical basis that the test measured a specific construct domain but, during administration in a real testing environment, it actually measures motivation to perform, it is no longer demonstrating construct validity.

Even the skills of other participants can affect scores. For example, if a test involves a team activity (e.g., four people lifting a piece of equipment into a truck), it could give inflated perceptions of one team member's strength. If the other members are strong and lift more than their share of the weight, it might appear that the fourth member is stronger than he or she really is. The potential sources of construct-irrelevant variance are essentially unlimited.

As these examples illustrate, validity is not an immutable property of a test. If something acts to alter test scores, it can affect the validity of selection decisions resulting from those scores. So ultimately, the goal is to validate the test *scores* that will be used in selection decisions. Because those scores can be affected by construct-irrelevant variance in ways that could differ from context to context, careful attention to validating the test scores in a way that emulates how the test will actually be used is important.

The test constructs involved in validation efforts are not the only constructs with which researchers need to be concerned. Measures of on-the-job outcomes (e.g., job performance, injuries, attrition) can also be affected. In a predictive validity study, ensuring that the outcome is measured properly is critical to drawing sound conclusions about the predictor (see the discussion on predictive validity in the next section). Construct irrelevance or deficiency in a validation study's outcome measure could lead the researcher to over- or underestimate a test's predictive validity.

Once a test construct is defined, the next decision is determining which type of validation study is most appropriate for supporting the test's use.

Content Validity

The process of establishing *content validity* involves soliciting expert judgment regarding the appropriateness of several aspects of a test's content, including the following:

- the extent to which a test covers the relevant content domain
- the extent to which the test's elements are proportionally representative of the domain
- the influence that construct-irrelevant variance can have on scores.

To address these, the content of the test could be compared with one of two possible construct domains:

- the construct that the test is supposed to measure (link 1 in Figure 5.1)
- the content on the job (link 4 in Figure 5.1).

For example, the content of a work-sample test of upper-body strength could be compared with the construct of "upper-body strength," or whatever other specific construct is believed to be involved in performing the tasks. To the extent that the content comparison (i.e., the content-validation process) supports the conclusion that there is good overlap between the test content and the construct, there would be support for link 1. The same work-sample test could also be compared with the domain of tasks on the job. To the extent that the content of the work sample shows good overlap with the domain of tasks on the job, there would be support for link 4. Both comparisons can provide important evidence for establishing the overall validity of a test as used for selection purposes.

In this section, we describe some important features to include in a content-validation study. However, it is worth noting that there are few agreed-upon guidelines for how such a study should be conducted (see Fitzpatrick, 1983, for other features that could be included). Ultimately, the decision is left to the researchers to determine those features that would best support the use of the test in their organizations. Regardless of which are chosen, justifications for each feature and the results of each step in the content-validation process described in this chapter should be documented in detail:

- Selecting SMEs. Careful attention to how SMEs are selected in a content-validation study is important. SMEs should be knowledgeable about the construct domain and about the job for which the test is being used. The use of multiple SMEs and a comparison of their judgments would be better than relying on a single SME or SMEs who all have the same type of expertise. The use of many SMEs with varying perspectives and expertise is ideal.
- Information that should be provided. The construct-validation process should provide SMEs with a definition of the construct being assessed and a set of clear guidelines for judging the content validity of the test, to include the factors that SMEs are expected to evaluate. Whenever possible, SMEs should be allowed to observe test administrations in a realistic testing setting and to take the test themselves.
- Factors that SMEs should evaluate. As noted previously, SMEs should evaluate whether the test is proportionally representative of a construct domain[23] and the extent to which it may be affected by construct irrelevance.[24] In making their judgments, SMEs should be instructed to consider the manner in which the test is administered. The following are examples of questions they should address: Does the test measure all aspects of the ability it is stated to measure? Could other factors (such as motivation, changes in instructions, encouragement by the administrator, or familiarity with the testing protocol or equipment) influence scores?

---

[23] Proportional representation has obvious application in the context of a multi-item test, such as a multiple-choice test. On such tests, it would make sense that the proportion of items covering one topic in the domain should be the same as the proportion of the domain that contains that topic and its criticality in overall job performance. How proportional representation applies to a physical test is not always obvious, but it is worth considering in judging the relevance of a test.

[24] There is no agreed-upon method for soliciting SMEs' judgment on either of these topics.

Content validity does have its limitations. The practice of establishing content validity is most often employed as the sole justification for a test's use when the test is a simulation of actual work activities. In that context, soliciting expert judgment to confirm that the simulation maps directly to important activities on the job, the physical demands are at the same level and intensity as those required on the job, and being able to perform the task is a necessary condition for being in the job is usually sufficient to justify the test's use. However, even for work samples, content validity alone may not suffice. For example, if a simulation requires physical skills that would be developed in training after selection has occurred, then applicants should not be expected to meet the physical requirements of the work-sample task prior to training. Moreover, applicants may differ in the extent to which they have already acquired the relevant skill. In this situation, justifying a simulation solely on the grounds of content overlap with the job could be easily criticized. Nevertheless, content overlap and the nature of the skill required to perform the simulation may be useful criteria for the previous step, choosing among alternative tests to be evaluated.

Criterion-Related Validity

Criterion-related validation involves measuring personnel on a selection test and examining the relationship between test scores and measures of important organizational outcomes (link 2 in Figure 5.1). This evidence can be collected in one of two forms: *predictive validity* evidence and *concurrent validity* evidence. The key difference between predictive validity and concurrent validity lies in when the selection test information is collected.

Predictive validity evidence requires longitudinal data, i.e., data collected on the same individuals at several different times. Predictor information (data on the selection tests) is collected on personnel at the time when the selection decisions will be made and then archived for future use. Those same individuals are then followed over time, and data on key organizational outcomes (e.g., injuries, job performance, attendance, training success) are collected after they have been on the job for some period of time. The outcome data are often collected weeks, months, or even years after the predictor information was collected.[25] Predictive validity is preferred over concurrent validity in the selection context because it can be designed to estimate the actual predictive results that would be obtained when the test is put into operational use.

In *concurrent validity*, the data on the predictors and outcomes are collected around the same time period. It typically involves collecting information about the outcomes of interest (e.g.,

[25] The appropriate time gap between collecting selection test scores and outcome measures is tied to the goal of the selection process. If the goal of selection is to predict long-term outcomes (e.g., long-term attrition from the job, likelihood of promotion), the time gap could span years. In other cases, the outcome could be weeks later (e.g., for graduation from a six-week training program).

injuries, job performance, attendance) on job incumbents and administering the selection tests to those same incumbents. Concurrent validation evidence is not ideal. Because both predictors and outcomes are collected simultaneously, construct-irrelevant variance associated with having been on the job cannot be ruled out. Experience, practice, maturation, and training are just some of the factors that could lead one to conclude that a test is a good predictor of key outcomes when, in fact, it is not. For example, in a concurrent-validity study, a work-sample test may distinguish those who are the best and the worst at those types of activities on the job. However, for applicants who have no experience on the job, the work sample may be unfamiliar. Those applicants may perform poorly on the test as applicants even though they would perform well on it later, after they have had training on the task or exposure to the job.

Predictive validity, in contrast, can be designed to avoid those concerns. If the data are collected in a way that emulates the timing of a test's anticipated use, predictive validation evidence is strongly preferred. The downside to predictive validity is that it can take longer to collect the necessary longitudinal data. If a concurrent-validation design is used to justify a test initially, an organization should (where possible) begin to collect longitudinal data to confirm the test's predictive validity after some period of time has passed.

Another factor that should be considered is whether the test itself or a related test has already been used to select people included in the validation sample. In other words, if all the personnel in a particular career field were required to demonstrate a high level of physical ability (e.g., strength) in order to qualify for training or for the job, then those people represent a restricted range of capabilities. In those cases, a predictive- or concurrent-validation study using that restricted sample would underestimate the relationship between the predictor and the outcomes of interest. There are statistical methods that can be applied to address this (for more information, see Sackett and Yang, 2000; however, if there is no variance in test scores in the group that is selected (e.g., test scores range from 1 to 10, but a 10 is required for entry into the job), then a criterion-related validity study cannot be performed. In those cases, creating what we refer to in this report as a *simulation study* may be a viable alternative.

In a simulation study, participants would complete the predictor test when selection decisions are made. However, instead of including only those who make it into the job, the study sample would include job applicants to ensure that the full range of scores is represented in the study.[26] The sample of applicants would then be trained on how to perform key job activities and, once trained, would be tested on a series of simulated job tasks (i.e., the simulated outcome measures). If a relationship were shown between the test and the simulated outcome, and job analysis data

[26] In cases in which participation in the outcome simulation might result in injuries for those with lower physical abilities, minimums for participation might need to be established. However, some range in test scores should be preserved because it is a fundamental necessity for estimating predictive validity. An alternative would be to reduce the physical demands in the outcome measure to allow participation by a larger group.

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.

One common criticism of a predictive-validity study is a failure to capture the appropriate organizational outcome (link 3 in Figure 5.1). A well-designed validation study outlines the types of outcomes that should be considered and documents why one outcome was chosen over others. The following are examples of the variety of outcomes that could be considered for use in a validation study, although some would be more easily justified than others for use in validating physical tests:

- training outcomes (training attrition, grades, instructor ratings, time to complete training, meeting specific course requirements)
- injuries (number, duration, severity, or medical costs of injuries to self or others in training or on the job; long-term injuries, such as repetitive-motion or overuse injuries; disability rates)
- job performance (supervisory ratings, peer ratings, or customer ratings of the quality of performance in on-the-job activities)
- productivity (number or speed of job activities accomplished)
- absenteeism (days missed)
- attrition from the job (e.g., attrition within one year)
- consequences (e.g., for rescue personnel, lives saved or lost; for maintenance personnel, equipment failures)
- promotions
- awards.

Which outcomes are best justified for supporting the use of a selection measure will depend on each organization's unique situation; however, some are more easily justified than others. In nearly all cases, job performance (i.e., how well someone performs important or frequent on-the-job tasks) is easily justifiable. In other cases, other outcomes may also be justified. For example, in the case of physically demanding jobs, training dollars lost to attrition, medical costs, or time lost due to injuries could be argued as important organizational outcomes to be predicted from a selection test.

Nevertheless, measures that might superficially appear justified may contain fatal flaws on closer inspection. For example, using training failure as an outcome assumes that the training content is vital to performance on the job. It further assumes that pass/fail decisions in training are well aligned with decisions about who will or will not fail on the job. If training outcomes are used in a validation study to support link 2, and evidence later indicates that training success or failure is not closely aligned with success or failure on the job, then the validation study results are fatally flawed. Collecting evidence for prediction of more than one outcome is always advisable. Similarly, job performance measures that are not construct-valid measures of the job performance domain (i.e., link 3 is not supported) could also lead to the conclusion that a validation study is flawed. In the context of Figure 5.1, for link 4 to be supported, link 2 and link 3 must be supported.

Convergent and Discriminant Validity

Convergent and *discriminant validation* evidence shows that the test correlates more strongly with measures of similar constructs (convergent) and less strongly with measures of different constructs (discriminant). Both are used to systematically rule out construct irrelevance and deficiency (providing evidence supporting link 1). For example, intelligence (i.e., aptitude) is one source of contamination (i.e., construct-irrelevant variance) that could be examined with a study of convergent and discriminant validity. In theory, a test of upper-body strength should correlate highly with other validated tests of upper-body strength, and it should not correlate highly with aptitudes that are conceptually different, such as intelligence. If a discriminant-validity study shows that a strength test is highly correlated with intelligence, then it is not a pure measure of the construct of strength.

In some circumstances, this type of contamination could be a serious concern. It is plausible that smarter people will figure out ways to perform better on the test. Maybe smarter people will read up on the test beforehand to learn the best techniques for performing well. Regardless of the explanation, in such cases, we would conclude that intelligence is adding irrelevant variance to test scores and, therefore, would have to question link 1.[27]

Showing *convergent validity* with another test already known to predict performance on the job can be a way to strengthen the argument for link 4 in the absence of a predictive-validity study. This approach would be particularly useful for finding less expensive selection tests as alternatives to those that are already known to predict organizational outcomes well. For example, if a test (such as a measure of VO_2 max) has been shown to be a good predictor of job performance but has other drawbacks (is determined to be prohibitively expensive and requires gender-based scoring), demonstrating convergent validity with a less expensive alternative measure (such as a timed one-mile run) could provide evidence supporting the use of the alternative measure. In such cases, the relationship between the two tests would be expected to be high (e.g., correlations of 0.80 or higher).

Discriminant validity is evidenced by results that a test does not correlate as highly with tests that claim to measure different constructs. For example, a test of upper-body strength should not correlate as highly with a test of lower-body strength as it does with a different measure of upper-body strength. How high the correlation between two measures of the same construct should be or how low the correlation between two measures of different constructs should be is

---

[27] Note that, if a physical ability test were found to be an impure measure of a physical ability domain but instead were contaminated by some other construct domain, such as intelligence or motivation, it could turn out to be an even better predictor of performance on the job because of the added contamination. The effect of this or any other type of contamination on a test's predictive validity, however, would need to be examined empirically to determine its effects to include whether it adds predictive bias. Some forms of contamination may reduce validity, and others may enhance it. Similarly, some may increase adverse impact, and some may reduce it. Regardless, irrelevant variance can change conclusions regarding the validity of test scores and, when identified, it warrants closer examination.

open to interpretation and should depend heavily on a sound theoretical understanding of the constructs in question. For example, most people who have developed strength in one part of their bodies have also developed strength in other parts of their bodies. In this way, we would expect a positive correlation between upper-body strength and lower-body strength.[28]

Convergent and discriminant validity can be used in employment settings and is discussed as such in *Standards* (Joint Committee on Standards for Educational and Psychological Testing, 1999), although it is most often applied to further theoretical understanding for the constructs measured by the test. Such theoretical understandings for the tests can be important, however, in defending the use of an employment test that might not stand up to criticism if it is the only method employed to justify a test's use. If it is the only evidence provided to support a test's use, the theoretical rationale linking the construct measured in the test to the job would need to be well thought out and strongly supported by existing evidence.

Fairness: Adverse Impact and Predictive Bias

As we discussed in Chapter 1, tests used for occupational screening should be fair. There are multiple dimensions to fairness (Joint Committee on Standards for Educational and Psychological Testing, 1999), but in practice it "is not simply a matter of whether or not test score averages differ by...[group], but whether or not there are differences in test score *predictions* by group" (Gebhardt and Baker, 2010b). If the predictions are equivalent (i.e., no differences in [estimated relationships between test scores and performance measures]), then there is no bias." *Adverse impact* and *predictive bias* are the two primary considerations for determining how a test affects relevant population subgroups. Opening occupations to women who can meet the job requirements focuses attention on gender, but the military also has a long-standing commitment to avoid unnecessarily restricting opportunities for other groups of service members.

Adverse impact occurs when one group's rate of selection is lower than that of another group.[29] For example, if 70 percent of male applicants and 40 percent of female applicants are selected, then the selection procedure has adverse impact against women. Adverse impact alone does not indicate that a test is unfair to the group affected. A test could show adverse impact for women, but it could still be a fair and accurate predictor of their ability to do the job. However,

---

[28] Convergent-validity and discriminant-validity estimates might be expected to differ in this example if the relationship is calculated by gender. If women tend to have greater lower-body strength than upper-body strength and the reverse is true for men, examining convergent validity without separating the results by gender could overestimate the strength of the relationship between upper-body and lower-body strength.

[29] In Title VII, adverse impact occurs when the selection ratio of one group is less than 80 percent of the selection ratio of another group. This is commonly referred to as the *80-percent rule*. The 80 percent rule does not apply to the selection of military personnel; however, similar principles regarding adverse impact are still applicable given that equal opportunity is strongly supported in the military.

the presence of adverse impact does indicate that close examination of a test's validity is needed to ensure that it is not also *biased* against that group. As an example, suppose that 80 percent of men and 30 percent of women meet the standard on a test for an occupation. This test is nevertheless valid if further examination confirms that these passing rates accurately reflect the proportion of men and women who can meet the physical requirements of the occupation.

Lay users of the terms often conflate *adverse impact* and *bias*, but, in personnel selection, these terms are not synonymous. In the personnel selection context, we are most concerned with a form of statistical bias known as *predictive bias*.[30] Predictive bias can take two forms. First, 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.

Second, it can occur when the predictive validity is equivalent for both groups but the test still underpredicts one group's performance relative to another group.[31] For example, if, for men, a score of 10 on a strength test suggests that they will fail and the same test is used for both men and women, then a 10 for women should have the same expected outcome—namely, failure. If, however, a study shows that a score of 10 would predict that women would, on average, succeed on the job when men with the same score would, on average, fail, the test would be under predicting female performance. Both types of bias need to be examined. If a test is discovered to exhibit either type of bias, it should not be used.

If a test's use is justified entirely on content validity or convergent and discriminant validation evidence, there are alternatives to examining predictive bias that could be applied instead. For example, SME review panels could be assembled to judge whether the test is biased against particular groups. One key element to attend to in those studies is the composition of the SME panels. For example, SME panels should include representation from members of the groups against which the test might be biased. See the *Standards* (Joint Committee on Standards for Educational and Psychological Testing, 1999) for more information about how to conduct SME panels for evaluating bias.

---

[30] Item bias is the other type of statistical bias that is defined in the *Standards*. Because most physical tests consist of only one item, examination of bias at the item level is not necessarily applicable. Regardless, bias in total test scores is the ultimate concern in the context of personnel selection. For more on both types of bias, see the *Standards* (Joint Committee on Standards for Educational and Psychological Testing, 1999).

[31] Underprediction could occur because of a problem with the test or with the construct being tested. If the test is not capturing the construct equally well for both sexes, the test is the problem. Alternatively, if, for example, men and women tend to use very different muscle groups to accomplish the same task on the job, the problem may lie in the choice of the predictor construct.

Additional Considerations in Collecting Validation Evidence

Collecting validation evidence is a complex process. The following are some additional guidelines for ensuring that an organization has strong validation evidence to support a test's use. We provide examples of other potential considerations in Table 5.1.

Document the Process

It is the validation researcher's duty to document all aspects of the research study design.[32] This includes the explanation for the processes used at each stage of the study and the results of those processes. The documentation should contain enough detail that another researcher could replicate the study. The researcher should also document the study's limitations and suggest follow-on research to address the limitations.

Apply Appropriate Statistical Methods

Criterion-related validity, convergent and discriminant validity, and adverse-impact and predictive-bias studies involve statistical analysis. The statistical methods for these studies require a careful design before data collection begins.

First, the study must have sufficient statistical power (i.e., a large enough number of test subjects) to obtain a precise estimate of the relationship between the test results and outcomes related to job performance. The power calculation should: (1) incorporate the best information available within the organization or from external sources on the expected distribution of performance scores on the tests and (2) be carried out for key population subgroups.

Second, the sample must be representative of the population in question, and the power calculations may show that some demographic populations (such as women or other groups) should be oversampled. In cases in which groups are oversampled, complex sampling statistics need to be applied in the subsequent analyses.

[32] An undocumented validation study is essentially the same as no study at all, if the details of the study cannot be recovered. If a test is challenged and a validation study was not documented, it cannot be used as justification for the test's use.

Table 5.1. Example Considerations in Validating and Selecting Tests

Consideration	Potential Resolution
Is the validation study considering the appropriate job outcomes?	A variety of outcomes could be selected for use in validating tests (e.g., training success, injury rates, performance on a work sample test, job performance ratings). Identify which outcomes the test should predict, and document the rationale for selecting specific outcomes. Consider conducting validation studies using a variety of important job outcomes.
Does the test leave out important physical skills needed on the job?	Examine predictive validity, content validity, and convergent and discriminant validity of the test.
Does the test measure something that is irrelevant to the job?	Examine predictive validity, content validity, and convergent and discriminant validity of the test.
Is the test biased against a relevant population subgroup (such as gender)?	Examine whether there are differences in the predictive validity of the test by group. Examine whether the test underpredicts performance of any group.
Would people improve on the test as a result of basic training or technical training?	Conduct a predictive-validity study estimating the amount of improvement expected for people at various score levels on the test. Collect selection test scores at the time when the test would be administered for screening during operational use. Measure again after completion of the training in question (e.g., after basic training). Create a score crosswalk to predict post-training scores. Use the crosswalk to evaluate applicants.
Can the characteristics required be trained? Could people easily develop the required physical skills through intensive practice and training?	Conduct a predictive-validity study estimating the impact, cost, and feasibility of a targeted training program. Include male and female participants at varying abilities; measure their abilities before entering the training and after completion of training. Examine the amount of change by gender, identify any injuries resulting from training, estimate the minimum start points associated with meeting the requirement by the end of the training, and identify total cost to train personnel to meet the requirements. If training does produce marked improvement without major injuries, adopt the training or publish the training program regimen and allow applicants to train on their own.
For jobs that have been closed to women, how can the performance of women be judged if they do not currently perform the job?	Establish a plan for how to examine this without injuring anyone or endangering the mission (if these are legitimate concerns). For example, an organization could test a sample of women, train them in key aspects of the job, and conduct work-sample simulations of the job to see how they would perform. Conduct the same training and simulations for a set of men who also have no experience with the job. Determine whether the predictive validity of the test is the same for both sexes. If the test is not valid for both sexes or underpredicts performance by women, look for a different test.
Is it easy to train for the test?	Examine whether training to increase scores on the test translates to increased performance on the job. If it does not, consider using a different test.

Third, the appropriate methods must be used to evaluate predictive bias and estimate the predictive-validity relationship as well as control for any confounding factors.[33] Which methods should be used depends on the statistical properties of the test scores and performance measures being evaluated. As just one example, the statistical methods employed to determine construct validity differ importantly based on the properties of the test data and performance data, including whether the data represent counts (e.g., number of test repetitions, such as pull-ups, in a given time period) a continuous measure (e.g., amount of time to complete a give number of repetitions). Similarly, the methods for determining validity differ when multiple tests are employed to assess a single physical capability or multiple performance measures are employed for the same job task. Therefore, we have not provided a summary of the relevant statistical methods for validation in this report. However, we note that considerable statistical expertise is required to ensure that a validation study is well designed and the tests selected based on the study results predict job performance with as much accuracy as possible and avoid bias toward any group of applicants.

Anticipate Potential Weaknesses in a Study's Methodology and Criticisms of the Results

No single study can address all possible criticisms. However, a carefully designed validation study will be subject to fewer criticisms than a poorly designed one. It is the researcher's responsibility to examine the methodology critically to identify flaws and weaknesses. When possible, weaknesses should be addressed through changes to the methodology. Any fatal flaws (i.e., factors that would cause the study findings to be useless) identified in the methodology should be remedied, and changes made should be documented. However, some reasonable criticisms will always remain. It is the researcher's responsibility to point those out and suggest additional research that can address them.

Collect Multiple Sources of Evidence

The ultimate goal of validation is to provide evidence that supports claims that scores on a test can be used for a specific purpose (e.g., that a timed one-mile run can be used to predict injuries on the job or job performance). Unfortunately, the best methods and strategies for accomplishing this cannot be laid out in a set of universal, predefined steps to guarantee success. Instead, the unique issues associated with that test and the intended use of the test should drive the choices made for selecting a validation approach.

---

[33] Personnel psychologists are typically well versed in the statistical techniques for estimating predictive bias and validation estimates; however, they may call on statisticians to assist them in dealing with unusual or complex statistical issues, such as oversampling. Most statisticians, in contrast, would likely not be familiar with the standard practices used by personnel psychologists to estimate bias and predictive validity, so they are not typically called on to conduct a validation study in its entirety.

For example, tests used for selecting or screening personnel are often validated using a criterion-related validity approach. Because the main purpose of selection is the prediction of future outcomes, criterion-related (and, more specifically, predictive) validity analysis is a sensible and intuitive source of evidence to support a test's use. This is a particularly useful approach for tests that have low fidelity to the job, for which establishing content validity could be challenging.

However, validation should not be conceived of as a singular event. A single study cannot address all of the potential concerns regarding a test's utility for selection. Instead, organizations should seek multiple sources of validation evidence whenever possible. Whether the multiple sources encompass different types of validation evidence (predictive, content, and convergent or discriminant) is far less relevant than whether those sources rule out different potential threats to validity.

Chapter Six. Establish Minimum Scores

Once a test or series of tests has been selected, the next step in the process is to establish the minimum scores that will reflect acceptable performance on the job.[34] The concept of "more is better" is not the relevant metric in establishing a minimum standard—despite the logic that better performers might be able to perform job tasks better. Rather, the goal in this step is to determine the minimum test score that corresponds to acceptable on-the-job performance. In this context, the Secretary's emphasis on not "reducing the qualifications for the job" is important for determining what level of performance should be considered acceptable.

Test score minimums for selection should be *criterion referenced* rather than *norm referenced.* This means that scores should be anchored to a concrete level of performance, such as lifting 80 pounds. They should not be based on a comparison to other performers, such as lifting as much as the top 60 percent of test takers. For example, if the on-the-job requirement is lifting 40-pound boxes, that requirement should be translated to a specific score on the predictor test. If, instead, the "minimum" score were defined by excluding the bottom 40 percent of the applicants, this approach could bias one group of applicants more than others—such as if women were less likely to meet the cut point than men—and would not be defensible.

Standard setting, or the process of establishing minimum cut scores, is distinct from validation. When used in employment contexts, it typically involves convening panels of experts to identify the test score that distinguishes a minimally competent performer from one who is not at least minimally competent. But because all experts may not agree, best practice requires a systematic approach that solicits the perspectives of a variety of people—referred to as a *standard-setting study.* The ultimate goal of standard setting is to make the resulting minimum cut score as objective and reliable as possible. Thus, documenting the process by which the minimum cut score is established is also critical.

There is no single approach to standard setting that would be justified in all cases. Instead, any of three general approaches could be applied, depending on the types of tests and data that are available. The first approach is to rely on data collected during the job. If this approach is not feasible, as is often the case, the second and third approaches involve conducting a standard-setting study to capture expert judgments of minimum performance on either the job or the test. These two policy-capturing approaches to standard setting are the ones that have received the most attention with respect to best practices.

[34] For a review of the practice of setting cut scores, see Cascio, Alexander, and Barrett, 1988.

Use of Job Analysis Data to Set the Minimum Score

It is possible to rely on job analysis data to justify a minimum score under certain conditions. This approach could be justified if all of the following are true:

- The test involves a high-fidelity simulation of key aspects of the job.
- The test shows good content overlap with the physical requirements of the job.
- Test scores are not expected to change prior to starting the job (e.g., if there is a time gap between testing and starting the job, and scores could change with intensive self-training or employer training, then job analysis information alone would not be sufficient to justify the minimums).
- The job analysis showed consensus across a representative sample of job incumbents or other SMEs regarding the minimum performance level that would be required to accomplish the task on which the simulation was based.

For example, if a job analysis shows good consensus among SMEs that dragging a body 50 feet is considered an important part of the job of a firefighter, and the test involves a simulation of dragging a 150-pound dummy 50 feet, then theoretically those who accomplish the task pass and those who fail to accomplish it do not pass. However, test minimums are sometimes not so straightforward.

Use of Expert Panels to Set the Minimum Score

Although the example above suggests that, with a well-designed job analysis, standard setting can be a simple and straightforward process, there are always elements of human judgment involved, and those elements could come under scrutiny. For example, some might argue that the dummy should weigh more than 150 pounds because many people weigh more than that. Or, they might argue that the distance should be less or more than 50 feet. In those cases, capturing experts' judgments on these issues becomes an important added step in supporting the minimum test standards.

In other cases, the test may not have enough fidelity to the job that a job analysis alone could be used to identify a cut score, or the job tasks may not have an obvious line distinguishing success from failure on the job. For example, for police officers, running to apprehend a subject may be an important part of the job, and one of their screening tests could include a timed one-mile run. But how fast should applicants be able to run a mile to be considered minimally competent at chasing down a suspect? People are likely to disagree on the answer. When the tests do not rely on simulated job tasks, when there is no obvious overlap between the test and the contents of the job, or when there is no obvious line distinguishing success from failure on the job, standard setting will require information other than the job analysis. In these cases, a policy-capturing standard-setting study to establish consensus on the job performance minimums is needed.

A policy-capturing standard-setting study can be approached in two ways. The first is to ask experts to identify a minimum level of required performance on the job, which can then be used to statistically estimate the minimum score required on the test. The second is to ask experts to identify a minimum level of performance on the test.

Capture Expert Judgments About Minimum Performance on the Job

In this approach, expert panels could be asked to judge at what level on the outcome measure a person has failed to meet the minimum requirements of the job. Then they could be asked to identify the consequences of false positives and false negatives and to determine what levels of false positives and false negatives are acceptable. The test scores that most closely approximate the acceptable levels of performance on the outcome measure, false-positive rates, and false-negative rates could then be established as the cut scores.

Statistically translating the minimum job performance levels established by the experts into a corresponding test score would require the following types of data elements:

- criterion-related validation data
- regression equations showing the formula for predicting an important aspect of performance on the job from scores on the test
- rates of false positives and false negatives associated with each score on the test.

This type of approach becomes increasingly difficult when the relationship between the test and the outcome is not strong. In those cases, or when no criterion-related validation evidence exists (e.g., validation was based entirely on content-validation evidence), the process will require additional SME judgment about scores on the selection tool itself.

Capture Expert Judgments About Minimum Performance on the Test

In this approach, SMEs are asked to identify the test score minimums that they believe distinguish between those who would be capable of performing on the job and those who would not. This is a bigger inferential leap. First, the SMEs have to draw conclusions about what constitutes minimum job performance (as described in the section above), and then they have to infer how the test relates to that minimum.

This type of standard-setting study is necessary if any of the following is true:

- The job analysis alone is insufficient to justify test minimums, and criterion-related validity data are not available.
- The criterion-related validity relationships are weak.
- Criterion-related validity data do not emulate the actual testing time frame, and no data exist to estimate the amount of improvement that could occur if someone worked on developing his or her skills during that time frame.

Although necessary in the above circumstances, employers might choose to pursue this type of standard-setting study even when criterion-related validity or job analysis data are available. Reasons for doing so could include

- verifying that the expert viewpoints are consistent with the results obtained using the other two methods
- ensuring buy-in from the SMEs and other stakeholders by involving them in the process
- showing that the cut scores have been endorsed by outside experts.

Methods for Obtaining Expert Judgments for Setting Standards

Approaches for standard setting have received a great deal of attention in educational testing and employment testing contexts. However, much of the published work focuses on multiple-choice tests of mental knowledge, skills, and abilities rather than physical aptitudes. Although the same general principles apply to mental and physical testing, techniques for establishing physical standards will, by necessity, differ from standards based on a multiple-choice test. For example, most well known methods[35] require that SMEs provide judgments about each item on a test. But many physical tests have only one score, the total test score. In the case of a test with only one score, a modified version of several well-known techniques could be applied.

One example is the contrasting-groups method. In this method, SMEs could be asked to sort people into two groups: those who are minimally competent on the job and those who are not. Using criterion-related validity data, test score distributions could then be created for each group (those judged as competent and not competent). The cut score could be set at the point at which the distributions overlap (to balance the rates of false positives to false negatives), or it could be set lower or higher to minimize the rates of either false positives or false negatives. The decision of how to balance the two types of selection errors should be made by consensus of the SMEs. For more on this and other common standard-setting methods, see Cizek (2001) and Livingston and Zieky (1982).

A Well-Designed Study

Unfortunately, there is no single correct method that is prescribed as best practice for conducting a standard-setting study, and research has shown that using different methods often produces different results. For that reason, we suggest using more than one method whenever possible to examine differences in results. Regardless of the method chosen, some key elements define a well-designed standard-setting study. Several of those elements are described below:[36]

- **Select appropriate SMEs.** This includes ensuring that they have sufficient experience with the job, are representative of the variety of personnel on the job, and represent a sufficient number of the key stakeholders. Examples of common considerations for selecting SMEs include representing different locations, levels of the job, levels of seniority, and races and genders. How many would be considered sufficient is largely

[35] The Angoff and Ebel methods are two examples (see Angoff, 1971, and Ebel, 1972).

[36] Hambleton (2001) provides a summary of many of these key features for setting standards in educational contexts, although many apply equally to the setting of standards in an employment context.

dependent on the context and number of stakeholders to be represented. For example, locations may differ in their requirements, so having a few representatives from a variety of locations would be ideal.

- **Select an appropriate methodology.** There are a variety of methods that have been established for setting standards, and many can be modified to apply to physical testing. As described in the section above, availability and quality of existing criterion-related validity and job analysis data should be one driving factor in determining the approach.
- **Establish consensus on the meaning of a *minimally qualified applicant*.** Most standard-setting studies ask SMEs to estimate the likelihood that a minimally qualified applicant would receive a passing score on an item or a test. This requires that the SMEs establish a common understanding for what constitutes being minimally qualified.
- **Match standard-setting goals to the purpose of the test.** If the test is designed to predict injury rates in training, SMEs should be asked to identify the score that is associated with the minimally acceptable likelihood of injury.
- **Evaluate reliability of the standards.** Collect data to estimate *interrater agreement* (i.e., how much individual raters agree or differ in their expert opinion) and *intergroup agreement* (i.e., the extent to which different groups of experts arrive at the similar or different conclusions after consensus). This would require a two-stage process, in which SMEs first establish minimums individually without discussion and then discuss the minimums with the group to arrive at consensus. To estimate group agreement, multiple groups would need to be included, and each group would need to arrive at consensus independently of the other groups. Lastly, when possible, researchers should attempt to replicate the minimum standards established by the study using an entirely different standard-setting method.
- **Orient SMEs to the test.** The results of the validation efforts should be provided in detail as part of the SMEs' introduction to the test. It is also common to ask SMEs to take the test to help them get a sense of the difficulty of the test. The purpose of the test should also be described in detail (e.g., it is being used to predict injury rates in training or ability to perform a critical task). SMEs should have a chance to ask questions about the test and about the elements of performance that it is designed to predict.
- **Use predictive-validity results to guide the standard-setting process.** Take as an example a predictive-validation study that shows that a score of 30 on a lifting task is associated with a 10-percent chance of injury in training, and a score of 29 is associated with a 12-percent chance of injury. If the SMEs determine that an 11-percent chance of injury is the highest they will accept, then the test minimum could be set at 30. The extent to which the predictive-validity test mimics the actual operational use of the test (including such factors as the amount of time between testing and the outcome of interest) can affect the appropriateness of this approach.
- **Provide SMEs with clear instructions and training on the standard-setting process.** This includes training them on the purpose and goals of the process, defining key terms, and explaining the materials to be used.
- **Exclude information regarding pass rates from the initial decisionmaking process.** This avoids both the perception and the reality that SMEs may be establishing minimums using quotas (i.e., a desired acceptance rate for a particular group) or norm-referenced scores rather than criterion-referenced scores. Overall pass rates can be considered later in the process, but only after a first pass at an SME consensus has taken place and its

results are documented. If, at that point, the pass rates have been set so high that too few would meet them, SMEs could be asked to reconsider their recommendations in light of the pass rates.[37]

- **Ask SMEs for feedback on the standard-setting process and the resulting standards they set.** Part of the goal of standard setting is to establish standards that those involved in and those external to the process will agree seem reasonable and well supported. If the SMEs do not believe that the final cut scores are appropriate or that the process used to arrive at them is flawed, the process should be reevaluated. Feedback could be collected systematically through a questionnaire administered at the end of the standard-setting process.
- **Document the entire process.** This includes documenting SME selection criteria, SME demographics, the information and instructions given to SMEs, the definition of minimally qualified, the results of the individual SME judgment process, the results of the group consensus process, comparison of results by location, and SME feedback on the appropriateness of the process.

We provide examples of these and other methodological considerations in Table 6.1.

[37] Note that many best-practice methods for standard setting recommend sharing pass rate information with SMEs at the start of the process. We suggest avoiding that because of concerns that physical standards may be set too low in an attempt to make accommodations for certain groups.

Table 6.1. Example Considerations in Establishing Minimum Test Scores

Consideration	Potential Resolution
Are test score minimums set too high? Will people be unfairly and unnecessarily excluded from the job?	Conduct a standard-setting study to establish test score minimums.
Are test minimums set too low? Are people being allowed into the occupation who cannot perform, are likely to injure themselves or others, or are unlikely to complete training?	Conduct a standard-setting study to establish test score minimums.
Are the standards based on someone's opinion, rather than scientific data?	Conduct a formal standard-setting study that uses systematic efforts to solicit expert judgments, evaluates the accuracy and reliability of those judgments, and documents the entire process. The more systematic and better designed the process (i.e., the more empirical it is), the more likely the results will be replicable.
Is there always someone else who can help do the physically demanding work, so the minimum should be adjusted to acknowledge that? Or, if a task is rarely performed but is critical, should the standard be set at a level that ensures that everyone can perform it?	In the information given to the experts during the standard-setting process, include job analysis data on frequency, importance, duration, and percentage of people performing the task.
Are the experts knowledgeable about the job or the requirements at all locations or under all circumstances?[a]	Chose experts carefully. Include representatives of all job locations, individuals with extensive experience in the job at the appropriate level (e.g., apprentice level), and enough experts that diversity of perspectives and experiences is adequately represented in the group.
If the standard-setting process were repeated with a different set of experts, would you have different results?	Include more than one panel of experts, and have each panel independently set standards. Compare the results to see whether there are differences.
Does requiring consensus on the standards mask important disagreements among experts?	Solicit individual perspectives on the standard prior to allowing any group discussions. Examine the variability in individual perspectives, and ensure that different perspectives are considered during group discussions.
Have you involved a diverse sample of job incumbents (e.g., women or underrepresented race/ethnicity groups)?	For example, if an insufficient number of female job incumbents are available (as would be the case in jobs previously closed to women), consider bringing in a panel of qualified women (e.g. experienced in other physically demanding occupations) to observe and learn about the job and involve them in the standard setting process. The same approach could be used with other groups underrepresented in the occupation.
Would using a different standard-setting technique produce different results?	Use more than one technique, and compare the results.
Are the experts capable of judging how score levels equate to performance on the job?	Ask experts to identify minimum levels of performance on the job. Check this against the minimums they establish on the test. Or establish a crosswalk that relates test scores to performance, and compare the job performance minimums to the corresponding test minimums.
Would some people currently working in the job not meet the minimums?	Administer the test to people on the job, determine who would not meet the minimums, and explore why they are still on the job. Is the test not sufficiently correlated with performance? Or are some people dropping below acceptable levels of job performance? If the latter, consider implementing annual testing to ensure that performance standards are being met. If the former, consider other tests that better approximate job tasks.

[a] For the military, personnel in some jobs may be knowledgeable only if they have been in combat or deployed during wartime.

Chapter Seven. Implement Screening

When has enough information been collected to support an organization's use of a screening tool? There is no clear answer. At a minimum, an organization should have formally documented the following:

- a clear statement of the intended uses for the test
- a detailed job analysis that supports the test's use for that purpose
- a summary of existing research literature on tests like the one to be implemented
- at least one solid validation study (more is always better) showing that the test is at least equally as valid as other reasonable options
- an examination of the test's adverse impact and evidence showing no consistent predictive bias against subgroups (e.g., by gender or other characteristic)
- a justification for selecting this test instead of other reasonable options. This is particularly important when the test shows adverse impact against subgroups.
- clear instructions for the proper test administration procedures and permitted uses of test scores. These should be consistent with the manner in which the test was validated.

Attending to key issues during the implementation step is vital to ensuring that the test is implemented in a manner that is consistent with the results of the validation and standard-setting efforts. In this chapter, we discuss a few key issues that should be addressed during implementation. We provide examples of these and other considerations in Table 8.1.

When the Test Should Be Administered

The timing of test administration is important. Tests that are administered far in advance of the work to be predicted should have evidence to show that the time gap does not change the validity of the test or the interpretation of test scores. For example, in some cases, scores collected before the time gap may underpredict or overpredict later performance. Overprediction could occur if applicants become complacent and reduce their physical activity while waiting to start their job assignments. Conversely, underprediction could occur if applicants increase their physical activity during that same time period. Basic training would be one event that would be expected to improve all applicants' physical abilities, resulting in systematic underprediction for everyone, unless training effects are accurately taken into account.

How much underprediction could be expected is a question best addressed by research. Some studies have examined improvement resulting from basic training (for a review, see Vickers and Barnard, 2010); however, amount of improvement is likely to be test dependent and to vary by the content of the training and an individual's physical ability levels at time of entry. For that reason, additional data collection on the operational test data may be needed to estimate the amount of underprediction that would occur in each operational circumstance. Ideally, this issue would be addressed empirically during the validation process. For example, a predictive-validity

study can be structured such that the selection test scores are collected at the time of selection and retesting is included after training or other relevant events have occurred. Relying solely on someone's expert judgment about the magnitude of the expected improvement is not recommended.

Regardless, minimum cut scores should be set at levels that allow for possible improvement and avoid underprediction. If test score minimums are lowered or if there is a potential that test scores will overpredict some people's performance, retesting after the intervening time period (e.g., just prior to job entry) should also take place to ensure that personnel are still capable of meeting the minimums required on the job.

Standardize the Test Administration Procedures

One element of fairness concerns each applicant having an equal opportunity to demonstrate his or her capability on the test.[38] Standardizing test procedures across administrations and locations is one way to ensure that.

The first steps to standardization are creating clear documentation of the proper administration procedures and ensuring that the equipment and testing environment is the same at all locations. The procedures, equipment, and testing environments that are established for the test's operational use should be consistent with the way in which the test was administered during validation. If it is not, the differences should be explained and justified and possible consequences for validity should be explored. Deviations from the protocol across locations or test administrations should be eliminated to ensure test fairness.

The next step involves training administrators to adhere to those procedures. All procedures should be clearly communicated to the personnel administering the test. They should receive training in those procedures and should be tested on them to be sure they are following the procedures correctly. Providing explanations about the importance of adherence to the procedures is important and may help ensure that administrators conform to them. Reduced predictive validity of the scores, lost training dollars, or poor job performance for false positives; perceptions of fairness for those not selected; and potential for legal action are some of the justifications that could be provided.

The last step involves conducting regular quality assurance checks to make sure that administrators are adhering to correct procedures, the testing environment is still comparable across locations, and the test equipment is still functioning appropriately.

---

[38] See the *Standards* (Joint Committee on Standards for Educational and Psychological Testing, 1999) for more information

Informing Applicants About the Test

A second element of test fairness is ensuring that all applicants have an equal opportunity to prepare for a test (see Joint Committee on Standards for Educational and Psychological Testing, 1999, for more information). To ensure that all applicants have an opportunity to prepare, they should be informed about the test as far in advance as possible. The following is the type of information that should be provided:

- a description of the test
- how the test will be used (e.g., to qualify people for a particular job)
- the minimums needed to qualify
- instructions for how to take the test, including that they should try their hardest regardless of how easy or hard the minimum is
- instructions for how to prepare for the test.

Consider Phasing the Test in Gradually

When a new test is instituted, the organization might want to phase the test in so that applicants have enough time to become familiar with the test and prepare for it. The first few administrations could be conducted without using the test for selection. This would not only allow test takers to become more familiar with the test but also allow the organization to identify any problems during administration that were not anticipated, such as equipment malfunctions, inconsistencies in test administration, or applicant confusion about the test procedures.

In addition, by gradually phasing in the test, an organization would have time to collect additional data on differences across groups (e.g., gender differences) and examine predictive validity in an operational setting.

Table 7.1. Example Considerations in Implementing Screening

Consideration	Potential Resolution
Is any of the test equipment broken or inadequate? Are scores collected using inadequate equipment invalid?	Conduct regular checks to ensure that safety instructions are up to date and equipment is working properly and calibrated to perform the same across locations. Fix or replace any problem equipment.
Is the test administered in the same way as it was validated, so that the validation evidence applies?	Establish standardized administration procedures that are consistent with those used during validation. When procedures differ from those used during validation, document a rationale for the change and whether that change is likely to affect the generalizability of the validity findings. If an impact is likely, consider conducting research to revalidate the measure using the new administration procedures.
Is the test administered the same way for everyone, so that all test scores have the same meaning?	Standardize test administration so every test is administered the same way for every person tested. Train people in the administration procedures. Conduct regular checks to confirm that those procedures are being adhered to. Fix any inconsistencies in administration that are identified during the checks.
Do all people know about the test? Those who know about it may have an unfair advantage because they can prepare for it.	Ensure that the test and consequences of performance are highly publicized. This includes making information readily available on the Internet and ensuring that recruiters give the same information about the test to everyone, including when it is administered, what it is used for, and ways to prepare for it.
Did something happen during testing that affected individuals' test scores such that scores do not reflect test takers' true abilities?	A variety of factors can interfere with testing that could result in inaccurate test scores, including equipment malfunctions, mistakes by the test administrator, performance anxiety, a test taker misunderstanding test instructions, or recent stressful life events. To ensure that individuals have a fair chance to demonstrate their abilities, there should be opportunities for at least one retest at a later date.
Is the test perceived as unfair? Is it clear to the test takers how the test relates to the job?	Provide test takers with information about why the test is used and information about the data that support its usefulness for predicting success on the job. Establish a process for handling and resolving test complaints. Also, see above issues regarding fairness.
Does some practice during test administration or misinformation about the test discourage members of certain groups from participating in testing or volunteering for the job?	Ensure that both men and women, and members of all groups, are equally aware, well in advance of testing, of the purpose of the test, test procedures, and how to prepare for the test. Conduct surveys or interviews with test takers to better understand differences in their perceptions of the test, and identify ways to correct perceptions of unfairness.

Chapter Eight. Confirm that the Tests Are Working as Intended

Once initial standards for entry into physically demanding occupations are established, they will need to be the subject of 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. Constantly changing technology and mission can significantly alter the requirements of the job. For all these reasons, the research effort should be treated as an ongoing process, one that should continue long after a test has been implemented.

Organizations should also revise testing policies as new research findings arise. Ideally, the organization's approach to these changes would be proactive—made in response to its own ongoing research—rather than in reaction to a challenge of the test's validity. To make sure changes are proactive, the organization should keep abreast of new developments in the field and continue to collect and analyze data to support a test's use. This chapter discusses examples of the types of proactive research and data-collection efforts that should be pursued. In Table 8.1, we provide examples of key considerations for confirming that test scores are working as intended.

Table 8.1. Example Considerations for Confirming that Tests Are Working as Intended

Consideration	Potential Resolution
Has the job has changed, or are the requirements outdated?	Conduct job analyses regularly (e.g., every three to five years) to determine whether there are meaningful changes in the job. For jobs in which change is occurring intentionally (e.g., two jobs are being merged into one), conduct a job analysis to identify the changes. Explore whether the changes should affect the types of tests that should be used or the minimum scores on existing tests. If so, conduct new validation or standard-setting studies to address the changes.
Are the results of some aspect of the process for establishing requirements in question?	Conduct additional research to address the element in question.
Have people started training to perform better on the test?	Monitor the type and amount of training individuals do to prepare for the test. If training may be interfering with the predictive validity of the scores, a new validation study is needed to determine the impact of that interference.

Institutionalize Research to Support Policy Changes

Ideally, several research efforts would be institutionalized as part of a regular operational data-collection activity for each occupation.

Reexamine Job Analyses

Job analyses should be redone on a regular basis to ensure that job requirements have not changed. For jobs that are not expected to change, a job analysis could take place every five years or so. However, for jobs in which the physical demands are constantly changing, more-frequent updates may be needed, and the organization could institutionalize a systematic process for identifying those fields that require closer examination. For example, if a career field's injury rates in training or on the job exceed some prespecified amount, one of those conditions might trigger a job analysis. A short annual workforce-wide survey inquiring about the physical requirements of the job could be developed to flag career fields that need to be examined more frequently.

Continuously Collect Longitudinal Predictive-Validation Evidence

Collecting and retaining data as part of normal operations would allow an organization to regularly conduct predictive-validity analyses and update them as needed. For example, predictor scores on tests that have been put into place for operational use should be collected and retained on applicants, as well as on people selected.

Reexamine Test Score Minimums

When a job analysis shows that a job has changed, new validation information and new standards should be established. Even when a job has not changed, a periodic reexamination of cut scores would still be warranted to show that the minimums are not outdated.

Reexamination would reveal whether an initially valid test stops being a useful predictor of performance because test takers start training specifically to score well on the test. This can occur when test takers prepare by developing only the narrow set of skills addressed on the test and not other related skills required on the job. If for example, pull-ups could be a good predictor of box-lifting capability on the job initially and if test takers start training and focus on improving pull-up proficiency only, pull-ups might predict their box-lifting performance on the job less well. Information on whether individuals are training for the test and whether the type and amount of training affects the predictive validity of their scores can be collected and used to determine if a test should be changed to better reflect the job's overall requirements rather than a narrow aspect of it.

Collect Test-Taker Reactions and Job-Incumbent Perceptions of the Tests

Regular collection of this information is useful in determining the continued effectiveness of tests.

Evaluate Whether Administration Procedures Are Being Followed

If a test is being administered according to the established guidelines and is still functioning properly, people should score similarly if they are retested. To assess whether tests are being administered properly, an organization could retest a representative sample of personnel on a regular basis. For comparison purposes, the retest should be completed under controlled conditions (e.g., at a new location and by someone known to use the proper administration procedures). Conducting regular field observations of test administration practices across sites is another way of ensuring consistency. Without these types of regular checks, there would be no way of knowing whether or not the data being collected and used for selection is accurate.

Conduct Additional Research as Needed

Although studies regarding most aspects of the validation process need to be repeated over time, some specific efforts will not need to be repeated. Unique research efforts, designed to address a specific concern, are just as important as recurring efforts for ensuring the validity of established standards. The following research questions are examples of ones that could be addressed by nonrecurring efforts:

- How much improvement can be obtained by additional training, and at what cost?
- Could the job be reengineered to reduce the physical demands?
- Are there new tests that the organization should consider adding to the *test battery* or using instead of the current tests?
- Do women do the job differently?
- What type of self-training would best prepare people to succeed on the test and on the job?
- Does the test still predict performance after an extended time period on the job?

Many of these research efforts are important to support test fairness and improve a test's utility. However, not all efforts need to be completed immediately. Having a cohesive plan for prioritizing the most-urgent efforts while still eventually tackling the other less pressing research issues would be the best way to ensure that resources are spent wisely.

Ongoing Personnel Research Efforts Are Not New

Creating institutionalized data-collection efforts and ongoing programs of research to support personnel policies is not new to the military. The Air Force has been collecting job analysis data on all enlisted career fields for the purposes of developing training protocols since the 1960s

(Mitchell and Driskill, 1995). The Army started collecting job analysis data for similar purposes in the 1970s (Brady, 2004). Similar job analysis efforts have been explored in the Navy and Marine Corps at one point or another (see Mitchell and Driskill, 1995 for a historical overview). Any existing systematic job analysis process in the services should be reviewed to determine whether it adequately addresses the physical requirements of the job or could be easily modified to do so. As we described previously, regular job analyses, along with systematic collection of test results and training and job performance measures, will help ensure that physical job requirements remain valid, fair, and supportable over time. Similarly, efforts undertaken by the services for the purposes of establishing physical job requirements could be designed to include elements for addressing other personnel issues as well.

The regular collection of these data would allow an organization to proactively evaluate and adjust current policies as needed. In addition, if a test were ever challenged, the availability of previously collected data would permit the organization to provide a swift, data-driven response supporting the way they are using a specific test.

Chapter Nine. Final Thoughts

The methods for establishing physical standards for specific occupations involve the six-stage process described in this report. The first four stages contribute to the initial development of the standards—the tests and minimum test scores that will be employed in screening for entry into an occupation:

- **Stage 1. Identify the physical demands of the job.** Define all tasks required on the job, and identify which of those tasks are physically demanding and which are not. Identify other relevant aspects of performance, such as injuries, that may be affected by physical ability.
- **Stage 2. Identify potential screening tests.** Explore past research on potential screening tests, articulate reasoned theories regarding the applicability of a particular tool, and identify varied options for inclusion in validation.
- **Stage 3. Validate the tests, and select those with highest validities and least adverse impact.** Administer a range of tests to job candidates, and examine the relationship between test scores and important outcomes on the job (e.g., job performance, injury rates, productivity). From the results of validation studies, identify the best predictors of performance, with the least adverse impact.
- **Stage 4. Establish minimum scores.** Apply a systematic process to identify minimum test scores that should be established for entry into or continuation in a job.

Each stage is 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. Setting the right level involves finding the minimum score on each test that differentiates individuals who are able to complete training and carry out the work from those who are not. Setting the standards too low will result in higher attrition rates in occupational training programs or subpar performance on the job. Setting standards too high unnecessarily limits the pool of individuals eligible to enter an occupation and denies opportunity to individuals who could be successful in an occupation.

Gender-neutral standards are set without regard to gender and reflect only the physical capabilities needed to perform the tasks associated with the occupation. However, to ensure that standards are not biased against women (or other groups)—that is, do not more frequently screen out women who could be successful in an occupation than men—the processes of validating tests and setting minimum test scores must be based on data collected from women as well as from men. When an occupation has been closed to women, the developers of standards should 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:

- **Stage 5. Implement screening.** Establish a systematic method of test administration. Train personnel in applying that method, and begin screening personnel using the test.
- **Stage 6. Regularly confirm that tests are working as intended.** Verify whether test administration in practice adheres to the guidelines that were established. Determine whether job requirements have changed. Examine whether coaching or test preparation activities have compromised the test's validity. Reexamine predictive validity and adverse impact of the test.

Without careful implementation and ongoing monitoring and updating, even well designed standards will fail to screen individuals appropriately if the testing is done improperly and as the occupational tasks and equipment change over time. Similarly, ensuring adequate performance requires establishing appropriate physical standards for job incumbents based on the tasks they are expected to carry out over their career.

We have provided an overview of the methods and data required to conduct each of the six stages for establishing standards for physically demanding jobs and identified key considerations at each stage. However, we have not addressed the many technical details involved. These details are determined based on the specific characteristics of the occupation, the environment, and any unique statistical needs or other issues encountered by the analysts. Carrying out the work requires expertise in a variety of domains, including industrial and organizational psychology, exercise physiology or a related field, psychometrics, and statistics. These experts rely on the expertise of 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.

Throughout the report, we have stressed the importance of documenting the methods used at each stage to develop, implement, monitor, and update occupation-specific physical standards. Documentation is essential to defending the appropriateness and unbiased nature of the standards. If the original developers fail to document their work, those who follow will find it difficult to know whether or when the standards have become outdated because of changes in the characteristics of the occupation or the applicants. The documentation should specify how each stage was carried out and record the important analytic results, including the following:

- list of physically demanding tasks
- list of tests considered and reasons for selecting among them
- procedures for validating the tests and setting minimum scores, including number of and selection criteria for test subjects and demographic makeup of participants, data-collection methods, and statistical analysis
- methods for training test administrators and ensuring that the tests are administered correctly and fairly over time
- ongoing procedures for establishing that the standards are working as intended and are updated when necessary.

Phase II of the RAND Study

This report documents the first major task in this project. The remainder of the project focused on reviewing the methods being used by the military services to set gender-neutral standards, as required to implement the recent decision to remove the ground combat exclusion rule for women. The report documenting the results of the second study uses the concepts presented here as an analytical framework for reviewing the work of the services and it provides a description of the services' overall approaches and more-specific methods for standard development.

Glossary of Terms

adverse impact. The effect of an employment practice that applies identical standards to members of all population groups but results in a selection ratio for one group that is less than 80 percent of the selection ratio for another group. This is commonly referred to as the *80-percent rule*. Adverse impact does not necessarily imply bias.

assessment. See *test*.

bias. "systematic error that differentially affects the performance of different groups of test takers"(*Standards*, 1999, p. 31). See *predictive bias*.

bona fide occupational qualification. A characteristic sought in a job applicant that is permissible even if it discriminates among members of certain groups because that characteristic is materially important to the performance of the job. An example might be a requirement of a cleaning person to be of the same sex as the occupants of a locker room for which that person will be responsible.

concurrent validity. *Criterion-related validity* evidence in which the predictors and outcomes data are collected around the same time period. It typically involves collecting information about the outcomes of interest (e.g., injuries, job performance, attendance) on job incumbents and administering the selection tests to those same incumbents. See *criterion-related validity.*

construct. The underlying concept or characteristic that a test is intended to measure.

construct contamination. See *construct irrelevant variance*.

construct deficient. A test is construct deficient if it fails to capture an important element of the construct domain (such as a test designed to measure overall strength that measures lower- but not upper-body strength).

construct irrelevant variance. Variance due to factors that affect test scores but are outside the construct domain (such as a test designed to measure overall strength that requires verbal acuity).

construct validation. The process of collecting multiple sources of research-based evidence to support a tests use. See also *content validity*, *convergent and discriminant validity*, *criterion-related validity*.

content validity. The degree to which a test adequately samples the domain of interest. See also *valid*.

convergent validity. Validity evidence showing that test scores correlate more strongly with measures of similar constructs relative to measures of different constructs.

criterion-referenced score. A score that is anchored to a specific and concrete level of performance, such as lifting 80 pounds. Contrast with *norm-referenced score*.

criterion-related validity. Evidence that test scores are correlated with measures of important organizational outcomes.

cut score. See *standard*.

differential validity. See *predictive bias*.
discriminant validity. Validity evidence showing that test scores correlate less strongly with measures of different constructs relative to measures of similar constructs.
disparate treatment. Any practice that overtly treats one group (i.e., a category based on gender or other characteristic) differently from how it treats another group.
face validity. The lay perceptions of a test's validity.
false negative. In the selection context, a candidate's test result indicating that he or she would not be capable of performing the task for which he or she is being tested, when in fact he or she would be capable of doing so.
false positive. In the selection context, a candidate's test result indicating that he or she would be capable of performing the task for which he or she is being tested, when in fact he or she would not be capable of doing so.
fidelity to the job. A measure of the degree to which a test task resembles the job task that it is meant to measure. The greater the fidelity, the more closely the test task resembles the job task. A high-fidelity test, such as a work sample, has obvious overlap with the job. A low-fidelity task, such as a grip-strength test to screen for a firefighting job, may be a valid predictor of success but is more abstracted from the actual task (in this example, actual firefighting).
gender-neutral standards. Standards that are the same regardless of gender and equally valid in predicting job performance for both sexes.
interrater agreement. The extent to which individual raters agree or differ in their judgment.
intergroup agreement. The extent to which different groups of raters arrive at similar or different conclusions.
job analysis. The process of establishing an accurate accounting of the tasks or activities that take place in a job.
measure. See *test*.
norm-referenced score. A score that is defined by a comparison with other test takers' performance, such as lifting as much weight as the top 60 percent of test takers. Contrast with *criterion-referenced score*.
occupational analysis. See *job analysis*.
occupation-specific entry standard. 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 for a job.
performance standard. Occupation-specific job requirements for satisfactory performance, for example as described in the *job analysis*.
personnel selection. See *selection*.
physical standard. See *standard*.
predictive bias. A form of statistical bias. Predictive bias 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 the test still underpredicts one group's performance relative to another group.

predictive validity. *Criterion-related validity* evidence that is 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. See *criterion-related validity*.

requirement. See *standard*.

screen. Evaluate the physical abilities of job applicants or incumbents as part of a *selection* process. May also refer to *selection.* Also, *screen out* means to exclude people from entering or continuing in a job.

selection. Any point at which decisions are made that may exclude people from entering or continuing in a job. This includes, but is not limited to, when people are selected for or assigned to specific jobs, when they wash out or wash back because of an inability to meet training standards, or when they are required to demonstrate competence on a training or professional certification test, maintenance of a competency to continue in his or her current job, or mastery of a new competency to continue or move up in the job.

selection test. See *test*.

simulation study. A validation study in which 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.

standard. Used interchangeably with the terms *cut score* and *requirement*, *standard* 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 of an occupation-specific entry standard.

standard setting study. A study designed to identify the minimum score required on a test for selection into a job or for certification of minimum proficiency.

task analysis. See *job analysis.*

test. Broadly refers to anything that might be used to exclude or disqualify someone from a job (also referred to as a *measure*, *tool*, or *assessment*).

test battery. A collection of tests administered as a group.

tool. See *test*.

validation. The process of measuring, quantifying, and collecting evidence to support the use of the test measure as such a predictor. See also *content validity*, *convergent and discriminant validity*, *criterion-related validity*.

validity. The degree to which test scores accurately measures what they are purported to measure. See also *construct validation*.

work analysis. See *job analysis*.

References

Anastasi, Anne, "Evolving Concepts of Test Validation," *Annual Review of Psychology*, Vol. 37, No. 1, February 1986, pp. 1–16.

Angoff, William H., "Scales, Norms, and Equivalent Scores," in Robert L. Thorndike, ed., *Educational Measurement*, 2nd ed., Washington, D.C.: American Council on Education, 1971, pp. 508–600.

Arvey, Richard D., Timothy E. Landon, Steven M. Nutting, and Scott E. Maxwell, "Development of Physical Ability Tests for Police Officers: A Construct Validation Approach," *Journal of Applied Psychology*, Vol. 77, No. 6, December 1992, pp. 996–1009.

Baker, Todd A., and Deborah L. Gebhardt, "The Assessment of Physical Capabilities in the Workplace," in Neal Schmitt, ed., *The Oxford Handbook of Personnel Assessment and Selection*, New York: Oxford University Press, 2012, pp. 277-298.

Binning, John F., and Gerald V. Barrett, "Validity of Personnel Decisions: A Conceptual Analysis of the Inferential and Evidential Bases," *Journal of Applied Psychology*, Vol. 74, No. 3, June 1989, pp. 478–494.

Brady, Elizabeth J., "Occupational Analysis in the United States Army: Past and Present," presented at the annual meeting of the International Military Testing Association, October 2004. As of March 14, 2013: http://www.internationalmta.org/Documents/2004/2004053P.pdf

Brannick, Michael T., Edward L. Levine, and Frederick P. Morgeson, *Job and Work Analysis: Methods, Research, and Applications for Human Resource Management*, 2nd ed., Los Angeles, Calif.: SAGE Publications, 2007.

Campion, Michael A., "Personnel Selection for Physically Demanding Jobs: Review and Recommendations," *Personnel Psychology*, Vol. 36, No. 3, 1983, pp. 527–550.

Cascio, Wayne F., Ralph A. Alexander, and Gerald V. Barrett, "Setting Cutoff Scores: Legal, Psychometric, and Professional Issues and Guidelines," *Personnel Psychology*, Vol. 41, No. 1, March 1988, pp. 1–24.

Christal, Raymond E., *The United States Air Force Occupational Research Project*, presented at the State-of-the-Art in Occupational Research and Development, Navy Personnel Research and Development Center, San Diego, Calif., AFHRL-TR-73-75, July 10–12, 1973; published 1974.

Cizek, Gregory J., ed., *Setting Performance Standards: Concepts, Methods, and Perspectives*, Mahwah, N.J.: L. Erlbaum, 2001.

Code of Federal Regulations, Title 5, Administrative personnel, Chapter I, Office of Personnel Management, Subchapter B, Part 300, Subpart A, Employment practices,

Section 300.103, Basic requirements. As of March 12, 2013: http://www.gpo.gov/fdsys/pkg/CFR-2012-title5-vol1/pdf/CFR-2012-title5-vol1-sec300-103.pdf

Code of Federal Regulations, Title 28, Judicial administration, Vol. 2, Chapter I, Department of Justice, Part 50, Statements of policy, Section 50.14, Guidelines on employee selection procedures. As of March 12, 2013: http://www.gpo.gov/fdsys/pkg/CFR-2010-title28-vol2/pdf/CFR-2010-title28-vol2-sec50-14.pdf

Code of Federal Regulations, Title 29, Labor, Part 1607, Uniform Guidelines on Employee Selection Procedures (1978). As of March 12, 2013: http://www.ecfr.gov/cgibin/textidx?c=ecfr&rgn=div5&view=text&node=29:4.1.4.1.8&idno=29

Code of Federal Regulations, Title 41, Public contracts and property management, Part 60-3, Uniform Guidelines on Employee Selection Procedures (1978)

Dempsey, Martin E., and Leon E. Panetta, "Elimination of the 1994 Direct Ground Combat Definition and Assignment Rule," memorandum for secretaries of the military departments, Acting Under Secretary of Defense for Personnel and Readiness, and chiefs of the military services, Washington, D.C., January 24, 2013. As of March 12, 2013: http://www.defense.gov/news/WISRJointMemo.pdf

DOL—*See* U.S. Department of Labor.

Ebel, Robert L., *Essentials of Educational Measurement*, 2nd ed., Englewood Cliffs, N.J.: Prentice-Hall, 1972.

Farr, James L., and Nancy Thomas Tippins, eds. Handbook of employee selection. Routledge, 2010.

Fine, Sidney A., and Maury Getkate, *Benchmark Tasks for Job Analysis: A Guide for Functional Job Analysis (FJA) Scales*, Mahwah, N.J.: L. Erlbaum Associates, 1995.

Fitzpatrick, Anne R., "The Meaning of Content Validity," *Applied Psychological Measurement*, Vol. 7, No. 1, January 1983, pp. 3–13.

Flanagan, John C., "The Critical Incident Technique," *Psychological Bulletin*, Vol. 51, No. 4, July 1954, pp. 327–358.

Fleishman, Edwin A., *The Structure and Measurement of Physical Fitness*, Englewood Cliffs, N.J.: Prentice-Hall, 1964.

Gael, Sidney, *The Job Analysis Handbook for Business, Industry, and Government*, New York: Wiley, 1988.

Gebhardt, Deborah L., "Establishing Performance Standards," in Stefan Harry Constable and Barbara Palmer, eds., *The Process of Physical Fitness Standards Development*, Wright-Patterson Air Force Base, Ohio: Human Systems Information Analysis Center, December 2000, pp. 179–199. As of March 11, 2013: http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA495349

Gebhardt, Deborah L., and Todd A. Baker, "Physical Performance," in John C. Scott and Douglas H. Reynolds, eds., *Handbook of Workplace Assessment*, Hoboken, N.J.: Wiley, 2010a, pp. 165–196.

———, "Physical Performance Tests," in James L. Farr and Nancy Thomas Tippins, eds., *Handbook of Employee Selection*, New York: Routledge, 2010b, pp. 277–298.

Hambleton, Ronald K., "Setting Performance Standards on Educational Assessments and Criteria for Evaluating the Process," in Gregory J. Cizek, ed., *Setting Performance Standards: Concepts, Methods, and Perspectives*, Mahwah, N.J.: L. Erlbaum, 2001, pp. 89–116.

Hogan, Joyce, "Structure of Physical Performance in Occupational Tasks," *Journal of Applied Psychology*, Vol. 76, No. 4, August 1991, pp. 495–507.

Hogan, Joyce C., George D. Ogden, Deborah L. Gebhardt, and Edwin A. Fleishman, *Methods for Evaluating the Physical and Effort Requirements of Navy Tasks: Metabolic, Performance, and Physical Ability Correlates of Perceived Efforts*, Washington, D.C.: Advanced Research Resources Organization, 1979.

Hogan, Joyce, and Ann M. Quigley, "Physical Standards for Employment and the Courts," *American Psychologist*, Vol. 41, No. 11, November 1986, pp. 1193–1217.

Industrial/Organizational Solutions, "Fleishman's Taxonomy of Human Abilities," White Paper 015, 2010. As of March 14, 2013: http://www.iosolutions.org/uploadedFiles/IOS/IO_Solutions/Research_and_Resources/Agency_Resources/White_Papers/Fleishman-white%20paper.pdf

Joint Committee on Standards for Educational and Psychological Testing, *Standards for Educational and Psychological Testing*, Washington, D.C.: American Educational Research Association, 1999.

Landy, Frank J., and Joseph Vasey, "Job Analysis: The Composition of SME Samples," *Personnel Psychology*, Vol. 44, No. 1, March 1991, pp. 27–50.

Livingston, Samuel A., and Michael J. Zieky, *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests*, Princeton, N.J.: Educational Testing Service, 1982.

McCormick, Ernest J., Paul R. Jeanneret, and Robert C. Mecham, "A Study of Job Characteristics and Job Dimensions as Based on the Position Analysis Questionnaire (PAQ)," *Journal of Applied Psychology*, Vol. 56, No. 4, August 1972, pp. 347–368.

Messick, Samuel, "Test Validity and the Ethics of Assessment," *American Psychologist*, Vol. 35, No. 11, November 1980, pp. 1012–1027.

———, "Meaning and Values in Test Validation: The Science and Ethics of Assessment," *Educational Researcher*, Vol. 18, No. 2, March 1989, pp. 5–11.

———, "Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning," *American Psychologist*, Vol. 50, No. 9, September 1995, pp. 741–749.

Miller, Laura L., Jennifer Kavanagh, Maria Lytell, Keith Jennings, and Craig Martin, *The Extent of Restrictions on the Service of Active-Component Military Women*, Santa Monica, Calif.: RAND Corporation, MG-1175-OSD, 2012. As of March 11, 2013: http://www.rand.org/pubs/monographs/MG1175.html

Mitchell, Jimmy L., and Walter E. Driskill, "Military Occupational Analysis: An Historical Overview," in Winston Bennett Jr. and Jimmy L. Mitchell, eds., *Military Occupational Analysis: Issues and Advances in Research and Application*, Brooks Air Force Base, Texas: Air Force Materiel Command, AL/HR-TR-1995-0131, August 1995, pp. 1–13. As of March 14, 2013: http://www.dtic.mil/dtic/tr/fulltext/u2/a303687.pdf

Myers, David C., Deborah L. Gebhardt, Carolyn E. Crump, and Edwin A. Fleishman, "The Dimensions of Human Physical Performance: Factor Analysis of Strength, Stamina, Flexibility, and Body Composition Measures," *Human Performance*, Vol. 6, No. 4, 1993, pp. 309–344.

Myers, David C., Deborah L. Gebhardt, and Edwin A. Fleishman, *Development of Physical Performance Standards for Army Jobs: The Job Analysis Methodology*, Washington, D.C.: Advanced Research Resources Organization, ARRO-3045-FR, 1980.

Panetta, Leon E. Statement on Women in Service as Delivered by Secretary of Defense, Lean Panetta, Pentagon Press Briefing Room, Thursday, January 24, 2013

Palmer, George J., and Ernest J. McCormick, "A Factor Analysis of Job Activities," *Journal of Applied Psychology*, Vol. 45, No. 5, 1961, pp. 289–294.

Principles—See Society for Industrial and Organizational Psychology, 2003.

Sackett, Paul R., and Hyuckseung Yang, "Correction for Range Restriction: An Expanded Typology," *Journal of Applied Psychology*, Vol. 85, No. 1, February 2000, pp. 112–118.

SecDef—*See* Secretary of Defense.

Secretary of Defense, "Direct Ground Combat Definition and Assignment Rule," memorandum for 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.

Sharkey, Brian J., and Paul O. Davis, *Hard Work: Defining Physical Work Performance Requirements*, Champaign, Ill.: Human Kinetics, 2008.

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

Standards—See Joint Committee on Standards for Educational and Psychological Testing.

Terpstra, David A., A. Amin Mohamed, and R. Bryan Kethley, "An Analysis of Federal Court Cases Involving Nine Selection Devices," *International Journal of Selection and Assessment*, Vol. 7, No. 1, March 1999, pp. 26–34.

U.S. Department of Defense, "Defense Department Rescinds Direct Combat Exclusion Rule; Services to Expand Integration of Women into Previously Restricted Occupations and

Units," news release, Washington, D.C., January 24, 2013. As of March 11, 2013: http://www.defense.gov/releases/release.aspx?releaseid=15784

U.S. Department of Labor, Employment and Training Administration, *Testing and Assessment: An Employer's Guide to Good Practices*, Washington, D.C., 1999. As of March 11, 2013: http://purl.access.gpo.gov/GPO/LPS53092

U.S. General Accounting Office, *Physically Demanding Jobs: Services Have Little Data on Ability of Personnel to Perform*, Washington, D.C., GAO/NSIAD-96-169, July 1996. As of March 11, 2013: http://purl.access.gpo.gov/GPO/LPS29700

———, *Gender Issues: Information to Assess Servicemembers' Perceptions of Gender Inequities Is Incomplete*, Washington, D.C., GAO/NSIAD-99-27, November 1998. As of March 11, 2013: http://purl.access.gpo.gov/GPO/LPS16712

U.S. House of Representatives, "National Defense Authorization Act for Fiscal Year 1994," House Bill 2401, November 10, 1993. As of March 11, 2013: http://thomas.loc.gov/cgi-bin/bdquery/z?d103:H.R.2401:

U.S. House of Representatives Committee on Armed Services, report on House Bill 2401, July 30, 1993.

Vickers, Ross R. Jr., and Amanda C. Barnard, *Effects of Physical Training in Military Populations: A Meta-Analytic Summary*, San Diego, Calif.: Naval Health Research Center, Report 11-17, October 25, 2010. As of March 11, 2013: http://www.dtic.mil/dtic/tr/fulltext/u2/a554490.pdf

Wu, Yiyang, "Scaling the Wall and Running the Mile: The Role of Physical-Selection Procedures in the Disparate Impact Narrative," *University of Pennsylvania Law Review*, Vol. 160, 2012, pp. 1195–1238.

Draft Research Report

Establishing Gender-Neutral Standards for Closed Occupations

A Review of the Services Efforts to Date

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

RAND National Defense Research Institute

RR-1340/2 -OSD
December 2015
Prepared for the Office of the Secretary of Defense

This document has not been formally reviewed or edited. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors. RAND® is a registered trademark.

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 (SecDef, 1994) which restricted assignments of women to occupational specialties or positions in or collocated with direct ground combat units (SecDef, 1994). After announcing the decision to eliminate the rule, the Fiscal Year 2014 National Defense Authorization Act directed establishment of gender-neutral and valid physical standards prior to opening any of the restricted occupations and jobs to women and gave the services until October 2015 to demonstrate that such standards were in place.

In response, the Office of the Under Secretary of Defense for Personnel and Readiness asked RAND for assistance in evaluating whether the work undertaken by the services would satisfy the NDAA requirements. To accomplish this, we 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. Our Phase I work was completed in 2013 and OUSD/P&R shared that report with the services. The second phase of the work involved a review and evaluation of the methods used by the military services to meet the requirement. The review began in 2013 (during the services early planning stages) and ended in April of 2015 (about 6 months prior to the 2015 deadline). This report provides the results of that review.

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. Comments on this project report are welcome and may be addressed to Chaitra Hardison at Chaitra_Hardison@rand.org.

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

Table of Contents

Preface ... iii
Figures ... viii
Tables ... ix
Summary ... x
Analytic Framework ... xi
Evaluating the Services Efforts to Develop Physical Standards ... xii
Army Combat Arms ... xiii
Army Special Operations Forces ... xiv
Marine Corps Combat Arms ... xiv
Marine Corps Special Forces ... xv
Navy Special Operations Forces ... xvi
Air Force Battlefield Airmen ... xvii
Conclusions ... xviii
Comparing Across the Service Efforts ... xviii
Unavoidable Limitations ... xx
Other Crosscutting Issues ... xx
Final Thoughts ... xxi
Abbreviations ... xxii
Chapter 1. Introduction ... 1
Setting the Stage ... 3
Organization of this Report ... 3
Chapter 2. Recommended Processes for Establishing Physical Standards ... 1
1. Identify Physical Demands ... 1
2. Identify Potential Screening Tests ... 3
3. Validate and Select Tests ... 3
4. Establish Minimum Scores ... 4
5. Implement Screening ... 4
6. Confirm Tests Are Working as Intended ... 5
Summary ... 5
Chapter 3. The Analytic Approach for Evaluating the Services' Efforts ... 7
Chapter 4. Army Combat Arms ... 11
Occupational Assignment and Screening in the Army ... 11
Overview of the Army's Validation Effort to Date ... 13
Identifying Physically Demanding Tasks (Our Stage 1) ... 14
Winnowing the Simulation Activities ... 16
Identifying Potential Predictor Tests (Our Stage 2) ... 18
Validating the Screening Tests (Our Stage 3) ... 19

Our Assessment of the Army's Approach 20
Chapter 5. Army Special Operations Forces 23
Occupational Assignment and Screening in USASOC 24
Army Process for Establishing Standards 28
Job Analysis (Our Stage 1) 28
Establishing the Link between the Selection Criteria and Job Competencies (Our Stage 3) 29
Our Evaluation of Both Stages 29
Chapter 6. Marine Corps Combat Arms 33
Occupational Assignment and Screening in the Marine Corps 33
Overview of the Marine Corps' Validation Efforts to Date 35
Methods Applied in the First Criterion-Validation Study 36
Methods Applied in the Ground Combat Element Integrated Task Force (GCEITF) Study 48
Conclusion 55
Chapter 7. Marine Corps Special Forces 57
Occupational Assignment and Screening 57
MARSOC process for establishing standards 58
Job Analysis 60
Test Validation 61
Setting Standards 62
Our Evaluation 62
Chapter 8. Navy Special Operations Forces 64
Occupational Assignment and Screening 64
SEAL Training 68
SWCC Training 69
Navy's Process for Validating SEAL and SWCC Selection Standards 70
Evidence from Existing Studies 70
Establishing Valid Gender-Neutral SEAL and SWCC Standards 74
Chapter 9. Air Force Battlefield Airmen 81
Occupational Assignment and Screening in the Air Force 81
The Strength Aptitude Test (SAT) 82
The Physical Ability and Stamina Test (PAST) Test 83
Training and Continuation Requirements 84
Establishing Occupational Entry Standards for Battlefield Airmen 86
Job Analysis 86
Criterion-Related Validation Study to Replace the PAST 87
Our Evaluation 91
Chapter 10. Conclusions 94
Comparing Across the Services Efforts 94
Operationalizing "Physical Screening" 94
Comparing Highly Similar Jobs Across Services 94
Establishing Occupation-Specific versus Combat Arms-Specific Standards 95
Unavoidable Limitations in What Can Be Completed Prior to Opening Positions 97

No Existing Female Applicants, Trainees, and Job Incumbents 97
Unforeseen Impacts of Implementation of Testing 98
Research Today May Fully Support Implementing the Standards, But Future Research May Still Show Changes Are Needed 98
Other Crosscutting Issues 99
Formal Documentation of All Aspects of the Work Is Needed 99
Process for Establishing Minimum Acceptable Scores Still Needs to Be Reviewed 99
The Implementation Step Still Needs to Be Investigated 100
Research Needs to Continue After the Standards Are Implemented 100
Final Thoughts 101
Appendix A. Terminology Used in Setting Physical Standards 103
The Personnel Research Community Has Established Definitions 103
Terms and Concepts Needing Greater Clarity 104
Screening, Selection and Standards 104
Tests, Scores and Measures 105
Occupation-Specific Standards Versus Health and Fitness Standards 106
Gender-Neutrality and Bias 107
Validation of Selection Practices 108
Summary 110
Appendix B. Physically Demanding Occupations Already Open to Women 111
Overview of the Services Efforts to Establish Physical Standards for Open Occupations 111
The Air Force's Physically Demanding Occupations 112
The Navy's Physically Demanding Occupations 113
The Navy's Process for Validating the EOD, Navy Diver and AIRR PST Standards 120
References 122

Figures

Figure 2.1. Six Stages in Developing Physical Standards xii
Figure 2.1. Six Stages in Developing Physical Standards 2
Figure 3.1. Physical Standards Development Process 8
Figure 4.1. Eligibility and Training Requirements for Army Closed Occupations 12
Figure 5.1 Army SOF Screening Lifecycle 26
Figure 6.1 Entry and Training Path for Marines Combat Arms Branches 34
Figure 6.2. Hypothesized Relationship Between Individual and Unit Attributes in the Ground Combat Element Integrated Task Force Study 49
Figure 7.1 Marines Screening Lifecycle 59
Figure 8.1. Eligibility and Training Requirements for Navy SEALs 66
Figure 8.2. Eligibility and Training Requirements for Navy SWCCs 67
Figure 9.1. Eligibility and Training Requirements for Enlisted Battlefield Airmen 85
Figure 9.2. Incrementally Higher Minimums Account for Improvement Gained From Training 90
Figure B.1 Navy EOD Screening Lifecycle 115
Figure B.2. Navy Diver Screening Lifecycle 115
Figure B.3 Navy AIRR Screening Lifecycle 115

Tables

Table S.1. Summary of Key Features of the Service Approaches xix
Table 4.1. Physical Tasks for Infantry (11B) 13
Table 4.2. Initial Task List Used in the Simulation Observation Study 15
Table 6.1. Marine Corps Ground-Combat Enlisted Occupations with Physically Demanding Tasks and Closed to Women 38
Table 6.2. Proxy Tasks for Physically Demanding Tasks in Marine Occupations Closed to Women 40
Table 6.3. Mission Events by Occupation for Ground Combat Element Integrated Task Force Study 51
Table 6.4. Performance Measures for Ground Combat Element Integrated Task Force Study ... 52
Table 8.1. Test Scores of SEAL and SWCC Graduates and Non-Graduates 73
Table 9.1. Physical Ability Stamina Test Minimums for Enlisted Jobs 83
Table 10.1. Summary of Key Features of the Service Approaches 96
Table B.1 Minimum PST Scores 114
Table B.2. Female AIRR Recruiting Goal 118

Summary

Although the role of women in the military has been gradually expanding since World War II, over much of this period, women have still been banned from serving in specialties and assignments that involve direct combat on the ground. However, on January 24, 2013, Secretary of Defense Leon Panetta and Chairman of the Joint Chiefs of Staff General Martin Dempsey announced the decision to rescind the 1994 ground combat exclusion policy and the intention to "integrate women into occupational fields to the maximum extent possible" as of January 2016 (U.S. Department of Defense, 2013). This change in policy potentially opened about 230,000 positions that had been previously closed to women.

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. The National Defense Authorization Act (NDAA) for 1994, section 543, 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 FY2015 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." These gender-neutral standards are to be developed, reviewed, and validated no later than September 2015, as specified in the FY2014 NDAA (sec 524).

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 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 established in the first phase of our research as a framework.

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. We are concerned with standards that are used to make selection decisions—that is, decisions made that may exclude people from entering or continuing in a job. *Gender-neutral standards* are based only on the physical capabilities required to perform the job, are the same for men and women, and should not differentially screen out a higher proportion of members of one gender who are,

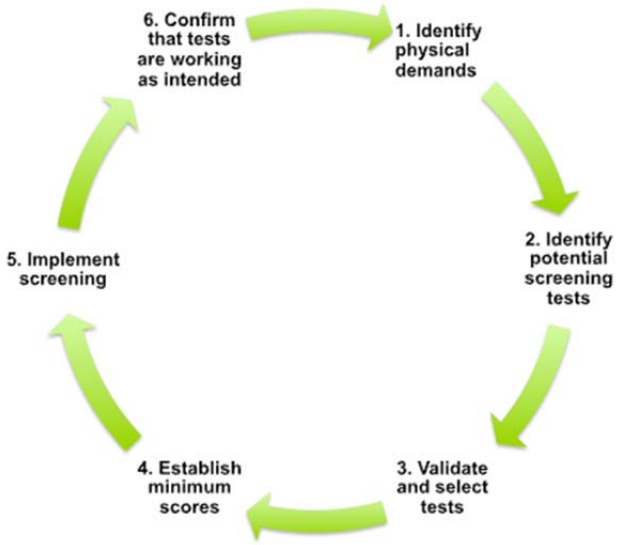
in fact, able to perform the job. Thus, the challenge for the military services is to identify a set of standards that is the same regardless of gender and valid in predicting job performance for both sexes.

Analytic Framework

To assist the military services in developing general and occupation-specific standards that are relevant to performance, we provided an overview of the processes recommended for developing those standards derived from 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 entering these careers. The recommended approach involves six stages. As shown in Figure S.1, the six stages are:

1. *Identify physical demands*. The process for establishing an accurate accounting of the 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 step in developing physical standards is to validate potential tests and identify those with the highest validity and least adverse impact. 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 step 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*. Once the previous steps have been completed and clear instructions for the proper test administration procedures devised, it is appropriate to begin using the screening tool in personnel selection.
6. *Confirm that tests are working as intended*. Once initial standards for entry into physically demanding occupations are established, they will need to be the subject of ongoing research to regularly confirm that tests are working as intended

Figure 2.1. Six Stages in Developing Physical Standards



We used the stages as a guide for evaluating the methodologies being used by the military services to set gender-neutral standards. Since the services are still in the process of developing standards, this evaluation focuses on the first three stages summarized in Figure 3.1 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 Services Efforts to Develop Physical Standards

In 2012, 21 percent of the Defense Department's active component authorizations were closed to women—just over 250,000 out of 1.2 million FY2011 authorizations (DoD, 2012). Since that time, a number of occupations have been opened to women. The remaining closed positions are not evenly distributed across the services. As a result, the magnitude of the challenges that the military services face as they put in place the elements necessary to open remaining closed positions to women differs 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 has 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 and therefore account for a smaller number of positions relative to the entire force. As with similar positions in the other services, these special operations positions in both the Air Force and Navy can be opened to women in 2016.

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

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 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 is roughly aligned with our recommended first stage (conducting a job analysis), whereas the second two steps most closely align with Stages 2 and 3 (selecting and validating selection tests chosen based on the data collected and the research literature). The first was aimed at defining and evaluating the critical physically demanding tasks in each specialty—which was conducted by first reviewing existing training activities, field manuals, and task lists to create a preliminary list of physically demanding tasks for each specialty and then revising the list through focus groups with subject matter experts and a survey of job incumbents.

The second effort involved administering realistic simulations of the critical physically demanding tasks to help identify candidate selection tests and develop a simplified set of simulations for inclusion in the criterion-related 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 subject-matter experts for use in the validation study. The third effort was a concurrent criterion-related validation data collection using the candidate selection tests and the simplified set of simulation activities in which the simulations were designed, measured, and analyzed with care and attention to detail. Approximately 150 participants were recruited for the criterion validation; researchers used regression analysis to determine the best predictor tests to include in the selection test battery.

The Army's work thus far generally aligns with recommended practices and the approaches used have many strengths. 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. A second strength was the amount of documentation available that provided sensible and understandable rationale for key study decisions. A third strength is the collection of information from multiple sources throughout the effort. Some gaps remain, however. One is the lack of examination of bias of the testing by gender; whether the tests predicted equally well for men and women is unknown. In addition, the concurrent-validation method they used establishes only the relationships when the predictor tests and simulations were collected at approximately the same time. It is not clear when the tests will be administered by the Army in practice, and this could have a significant impact 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 to address the establishment of score minimums for entry into the occupations.

Army Special Operations Forces

U.S. Army Special Forces Command (USASOC) regularly reevaluates training standards for special operations forces but initiated a new effort to validate these standards in response to the congressional requirement and enlisted assistance from the Office of Personnel Management (OPM) and the Naval Health Research Center. OPM was asked to conduct a new in-depth job analysis for the Special Forces and Ranger occupations to determine the knowledge, skills, and abilities (KSAs) required in these occupations; the job analysis is based on reviewing background occupational information, as well as conducting site visits and administering a survey. OPM is also exploring whether personnel assessments and standards are based on competencies required for that position. According to USASOC, OPM would use statistical techniques to determine the degree to which test activities during the current 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 appears that USASOC is 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. Although we were not able to evaluate important details of this process, OPM has a long history of job analysis and validation of selection practices that are generally consistent with recommended practices—so their approach is likely to be defensible. That said, some questions remain. Selection into the training programs is made using applicant rankings by senior USASOC personnel. The applicants selected have standard physical fitness test scores well above the minimums required to submit an application. It is not clear how the job analysis information could be used to assess the applicant ranking process or determine which training tests are most valid, what the minimums on the tests should be, and whether the tests are biased against any relevant groups. Also, additional evidence showing the link between the information collected in the job analysis and the screening and training criteria is needed (our recommended Stage 3). Moreover, the OPM job analysis as it was described to us does 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 Step 2 is being addressed by USASOC's current approach.

Marine Corps Combat Arms

The Marine Corps is relying on the results of two major studies in 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 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 the Marine Corps' existing Physical Fitness Test (PFT) and Combat Fitness Test (CFT) elements with performance in proxy tasks of the most physically demanding tasks identified in the first step. The Navy Health Research Center then analyzed the data after it was collected by the Marine Corps. Based on the findings the Navy researchers recommended a set of screening tests and minimum qualifying scores for selection into these occupations.

Although the process generally aligned with our recommended stages for developing standards, we identified several limitations in the data and the analyses that could affect the soundness of the findings. For example, although the participants included both men and women, nearly all of the men successfully completed all of the proxy tasks. As a result, the variance for the female participants is driving the relationship between test scores and the proxy performance tasks. With both the male and female data combined in the same statistical analyses without controlling for gender, the results can be misleading. Also, the decision to rely on existing fitness tests was made before the physical job tasks were identified and it is questionable whether the simulations can actually serve as realistic proxies for those physical job tasks. Moreover, the findings from the proxy simulations were intended to apply to all closed occupations regardless of whether that 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 an "integrated task force" to evaluate the performance of gender-integrated ground-combat teams—the Ground Combat Element Integrated Task Force (GCEITF). The task force consisted of 376 Marine volunteers, including 77 women, who were evaluated as they rotated through a series of simulated elements. This study was in progress at the time of our evaluation. The GCIETF could provide additional data and analysis that may address 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. However, 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 their 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 cannot fully evaluate how useful the experiment will be for this purpose.

Marine Corps Special Forces

The Marine Corps Special Operations Command (MARSOC) outlines three principle steps in their approach to establishing standards: (1) conduct a detailed job analysis for the special operations positions of interest that is focused on identifying the tasks and abilities required on the job; (2) validate standards in the Individual Training Course and Special Operations Training Course; and (3) validate the assessment and selection course standards, which includes

identifying selection factors and screening tests, 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 (though the Marine Corps has acknowledged that there might not be time to complete criterion-validation work prior to the deadline). Similar to the Army Special Forces, these efforts are solely directed at validating the selection that occurs *during* the training courses. At the time our data collection ended, no plans were in place to validate the processes used to screen people prior to entering training, which rely on rankings by senior MARSOC personnel of applicant packages.

MARSOC has contracted with the Office of Personnel Management (OPM) to execute their validation plan. Because of the timing of OPM's contract initiation, our evaluation is based on the scope of work OPM provided to MARSOC. Their description of the planned job analysis process is generally consistent with recommended practice; however, close examination of the documentation of the results of that work will still be important. Other steps beyond the job analysis, however, are laid out in less detail, making it difficult to judge whether the results will provide sufficient support for MARSOC's selection process. Details on the tests that would be identified to validate, 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. As a result of the lack of available documentation, there are large gaps in our understanding of the work that OPM is doing for MARSOC. When OPM provides documentation of the entire process and the findings, some or all of those gaps can likely be filled.

Navy Special Operations Forces

In the Navy, the Special Warfare Operator (SEAL) and Special Warfare Combatant-Craft Crewmen (SWCC, also known as Special Warfare Boat Operators) occupations are currently closed to women. The Navy's data collection efforts for validating selection standards does not include a re-examination of the physical testing requirements for screening candidates into training or for continuation in training; instead they are relying on previous research, a small portion 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 scores—not the least of which is the need to provide evidence that the scores would predict equally well for both male and female applicants, something past studies do not explore.

Instead, the Navy's data collection focuses on investigating the extent to which SEAL and SWCC selection requirements that occur *during* training are related to occupational performance. To update an older job analysis, the Navy relied on the input of subject-matter experts 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 our analytic

framework and an approach that would be strengthened if outside experts provided an independent assessment that agreed with the conclusions.

The Navy also included survey questions asking job incumbents to provide judgments about whether physical training activities during Hell Week were important preparation for mission success (an identical approach was use to validate the SWCC equivalent, called The Tour). Based on the results of these surveys, the Navy has concluded that Hell Week is valid preparation. However, job incumbent judgment alone provides a limited basis for such a claim. Given that some have expressed concerns over the years about the training difficulty differing from class to class and student to student, stronger evidence may be needed to refute claims such as these. In addition, collecting other more empirical evidence to support the link between Hell Week and actual on-the-job performance would go a long way towards strengthening support for the continued use of Hell Week as it now stands.

Moreover, none of the data collected by NHRC included females because there are no women currently on the job or in training, so it is unclear if the training-performance relationships would be the same for women and men. This is an area that should be explored further in the future. Our evaluation of the Navy's effort was based largely on a draft write-up of the intended methodology; once full documentation of the effort is available, some of the potential gaps we identified above may be eliminated.

Air Force Battlefield Airmen

In the Air Force, only seven occupations—as well as the associated units and training courses—are 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 subject-matter experts 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 a range of physically demanding job performance simulations. The results of this validation study will be used to establish the recommended annual testing standards for the battlefield airman operators and to establish the training entry requirements. Once the operator tests are selected and minimum test standards are set, the researchers plan 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 will 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-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 they plan to have data supporting many of the important links that are critical in a well-designed criterion-validation study. However, the formal write-up of the methods, analyses, and findings are still forthcoming and therefore many of the details of their data analysis decisions are still unknown to us and were not included in our evaluation. In addition, although there are many strengths to the approach that can lend credibility and support to any resulting test score minimums, there are some potential gaps in the work. In particular, examination of bias of the testing by gender is one area that was not addressed in the plans described to us.

Conclusions

Comparing Across the Service Efforts

Each service took a slightly different approach to amassing evidence to develop and support their screening standards. Differences in their approaches should not be taken to 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. We highlight several notable differences here.

- *Operationalizing "physical screening*." Each service conceives of their 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 and the Air Force's efforts for its special operations occupations will be used 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. However, in each case, the information obtained through the research is useful for informing the validity of the other screening elements.
- *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 across the Army and Marines efforts. The two services have taken very different approaches to establishing the standards 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 will likely need to be reconciled with attention to the 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 will be called upon 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. This too is an area where differences will be apparent and warrant justification.

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

Service	Selection Process Being Validated	Step 1 Job Analysis	Step 2 Identifying Screening Criteria	Step 3 Validation
Army Combat Arms	Screening before training	Review of existing job-analysis materials through 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 OPM using occupational information, site visits, job incumbent survey	Current training activities	Content validity, details to be determined
Marine Corps Combat Arms (phase-1 study)	Screening before training	Job tasks identified from current training and readiness manuals, which rely on occupation-specific task lists regularly updated based on SME review and a job incumbent survey	Elements of current Physical Fitness Test and Combat Fitness Test	Concurrent criterion-related validation to determine how well candidate tests predict performance on basic physical tests roughly similar to physically demanding job tasks
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 Forces	Training	New in-depth job analysis by OPM using occupational information, site visits, job incumbent survey	Current 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	Current 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
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

Unavoidable Limitations

No single research effort can address all issues, and no research study is without weaknesses and gaps. As a result, Stage 6, continued research, is an important next step after the standards are in place. Three gaps in particular are issues common to all of the services' work in support of standards for the closed occupations.

- *No existing female applicants, trainees, and job incumbents*. No women are in the closed jobs so there was not a pool of incumbent women for the researchers to draw upon as participants in the research with the same experiences or training as their male counterparts. This was an unavoidable dilemma. As a result, we strongly 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.
- *Unforeseen impacts of implementation of testing*. Implementation of the testing (Stage 5) itself can lead to unforeseen changes in the validity of the testing. This could apply regardless of gender, and it is something the services will need to watch closely. Collecting data on this in the months and years after establishing the standards will be important for ensuring that the tests and criteria perform as expected.
- *Future research may show needed changes*. The services' research efforts are intended to establish standards on the basis of the evidence amassed so far, but more research ultimately will be needed to fully determine whether and how well the tests and test minimums are working (Stage 6). We expect that as additional information is amassed and the available tests evolve, the services will need to make adjustments and refinements to the selection processes—a normal and necessary part of this process.

Other Crosscutting Issues

- *Formal documentation of all aspects of the work is needed*. Details such as the overall statistical and methodological approaches, summary statistics, data analyses, sampling approach and participant characteristics, etc. are all necessary for experts to be able to judge the soundness of the research findings. Without those details, evidence to refute any challenge to the selection practices ceases to exist. For that reason, we recommend that the documentation the services create include detailed write-ups of all research conducted to support and evaluate occupational physical standards—and that the services archive this documentation and make it available.
- *Process for establishing minimum acceptable scores still needs to be reviewed.* At the point of completing our research, none of the services had established minimum selection standards. However, this step is key to determining whether the standards are set appropriately so that the services are admitting people who are capable of performing on the job while not excluding valid candidates.
- *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 have occurred that could result in reduced validity.

- *Research needs to continue after the standards are implemented*. Not all research can be done a priori. More research will be needed over time. It will be important to follow up after implementing the standards to see if the standards have good predictive validity in practice.

Final Thoughts

The call to develop valid standards has been taken very seriously by the services. All of the services have dedicated a large amount of time and resources to their work in response to the lifting of the ground combat exclusion policy. As a result, the service efforts have been very large undertakings. Some 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. All have sought to involve personnel with the appropriate research background and expertise. Some services had the requisite experts in house, whereas others sought out the assistance of experts outside of their organization. The numbers of voluntary participants joining in the work have also been impressive. All told, the work that the services have put forth reflects a valiant effort to accomplish exactly what was being requested: the establishment of gender-neutral valid physical standards.

Abbreviations

AETC/A1	Air Education Training Command
AFQT	Armed Forces Qualifying 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
CRO	Combat rescue officer
CSO	Critical skills operators
C-SORT	Computerized Special Operations Resilience Test
DGCAR	Direct Ground Combat Assignment Rule
EOD	Explosive ordinance disposal
GCEITF	Ground Combat Element Integrated Task Force
HPP	Human Performance Program
IRB	Institutional Review Board
IST+	Initial Strength Test
ITC	Individual Training Course
KSA	Knowledge, skills, and abilities

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
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
PJ	Pararescue
POI	Programs of instruction
PST	Physical screening test
RASP	Ranger Assessment and Selection Program
SAT	Strength aptitude test
SEAL	Sea, air, and land
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 Forces Command
WARCOM	Warfare Command
WMD	Weapons of mass destruction

Chapter 1. Introduction

The role of women in the 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 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" (CRS 2013).

The change in policy meant that women could be assigned to any position for which they were qualified, with exception of "those units below the brigade level whose primary mission is to engage in direct combat on the ground."[1] Though 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 including physically demanding tasks the "vast majority" of women cannot do (SecDef, 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 FY2011 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.[2] Then, in 2012, DoD rescinded the co-location restriction, opening 14,000

---

[1] Les Aspin, Memorandum, Subject: Direct Ground Combat Definition and Assignment Rule. January 13, 1994.

[2] U.S. Department of Defense, *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.: Office of the Under Secretary of Defense for Personnel and Readiness, February 2012.

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 and Chairman of the Joint Chiefs of Staff General Martin Dempsey announced the decision to rescind the 1994 ground combat exclusion and the intention to "integrate women into occupational fields to the maximum extent possible" as of January 2016 (U.S. Department of Defense, 2013). This change in policy potentially opened about 230,000 positions that had been previously closed to women. In announcing the decision to eliminate the rule, the Secretary 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.

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. The National Defense Authorization Act (NDAA) for 1994, section 543, 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 FY2015 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." These gender-neutral standards are to be developed, reviewed, and validated no later than September 2015, as specified in the FY2014 NDAA (sec 524). 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 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 established in the first phase of our research as a framework. The report focuses on physical standards for military occupations that are closed to women. Appendix B addresses physical standards for physically demanding occupations that are open to women.

Throughout 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. We are concerned with standards that are used to make selection decisions—that is, decisions made that may exclude people from entering or continuing in a job. *Gender-neutral standards* are based only on the physical capabilities required to perform the job, are the same for men and women, and should not differentially screen out a higher proportion of members of one gender who are, in fact, able to perform the job. Thus, the challenge for the military services is to identify a set of standards that is the same regardless of gender and valid in predicting job performance for both sexes.

Setting the Stage

In 2012, 21 percent of the department's active component authorizations were closed to women—just over 250,000 out of 1.2 million FY2011 authorizations (DoD 2012). Since that time, a number of occupations have been opened to women. The remaining closed positions are not evenly distributed across the services. As a result, the magnitude of the challenges that the military services face as they put in place the elements necessary to open remaining closed positions to women differs 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 has 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 and therefore account for a smaller number of positions relative to the entire force. As with similar positions in the other services, these special operations positions in both the Air Force and Navy can be opened to women in 2016.

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 to evaluate the work being done by the military services to set gender-neutral standards. Chapter 3 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 forces. Finally, Chapter Nine discusses battlefield airmen—the Air Force's special operations positions that will be newly open to women—as well as efforts to validate the strength aptitude test for use in setting standards in other physically demanding jobs. The report concludes with 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 included two appendices as supplemental information. Appendix A provides an overview of many of the technical terms that are 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 that are already open to women.

Chapter 2. Recommended Processes for Establishing Physical Standards

Civilian employer's 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 general and occupation-specific standards that are relevant to performance, we provided an overview of the processes recommended for developing those standards.[3] These were grouped into six stages to reflect the fact that the process necessarily involves attending carefully to each stage in the process. For each stage, important features associated with good practice in addressing the step were described. We based these 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 entering 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 organization psychology, exercise physiology or a related field, psychometrics, and statistics. These technical experts also rely on the expertise of subject matter experts 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 step along with careful documentation of the actions taken is central to developing defensible physical standards. An overview of the six steps is provided below.

In addition, for further reference, we have provided an appendix (Appendix A) that explains many of the key terms used in this and later chapters in the report.

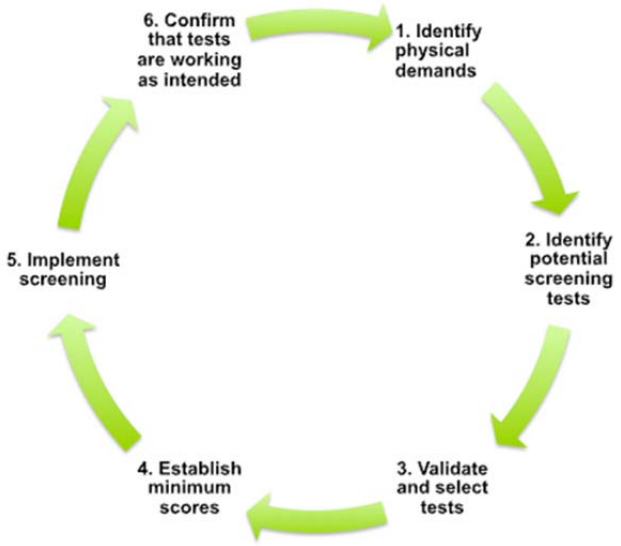
1. Identify Physical Demands

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

[3] See our companion report RR 1340/1-OSD.

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.

Figure 2.1. Six Stages in Developing Physical Standards



In establishing gender-neutral requirements for entry into physically demanding jobs, the focus is on applicant selection and the job analysis will be used to design an appropriate selection system. So the job analysis should identify and describe in detail the physically demanding tasks the applicants would need 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 subject matter experts 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. Similar issues arise in setting standards for continuing in a job, but the test subjects would include 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.

2. Identify Potential Screening Tests

Identifying potential tests that might be used to screen job applicants (or job incumbents) is the next step in developing physical standards. In this context, we use screening to refer to evaluation of individuals' physical skills relevant for performing the job tasks described in Step 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 may 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 are 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 the ability to accurately replicate 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. Further, 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 may be a more feasible approach.

Where physical standards already exist for the occupation, the test(s) used will 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 can also be considered.

3. Validate and Select Tests

The third step 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 is to provide evidence that the selection test predicts important outcomes on the job.

Best practice requires that 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 may be selected who are not capable of performing on the job or candidates may 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 expertise and a careful design before data collection begins to ensure results are as accurate as possible and avoid bias toward any group of applicants. Finally, organizations should seek multiple sources of validation evidence whenever possible.

4. Establish Minimum Scores

The next step in the process is to establish the minimum scores that will reflect acceptable performance on the job. The goal in this step 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, 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 the Secretary's commitment to not "reducing the qualifications for the job."

The process of establishing minimum cutoff scores, referred to as standard setting, is distinct from validation. When used in 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 may 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.

5. Implement Screening

Once the previous steps have been completed and clear instructions for the proper test administration procedures devised, it is appropriate to begin using 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 that is consistent with the results of the validation and standard-setting efforts.

The timing of test administration can influence results. Tests that are administered far in advance of the work to be predicted should have evidence 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 under predict 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.

6. Confirm Tests Are Working as Intended

Once initial standards for entry into physically demanding occupations are established, they will need to be the subject of 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 mission can significantly alter the requirements of the job. And new research findings may arise 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 that is not new to the military services.

Summary

The methods for establishing physical standards for specific occupations involve the six-stage process described. 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 as well as from 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 if the testing is done improperly or as occupational tasks and equipment change over time.

Chapter 3. The Analytic Approach for Evaluating the Services' Efforts

We used the stages and the best-practice methods described in the previous chapter as a guide for evaluating the methodologies being used by the military services to set gender-neutral standards—and which are described in the chapters to follow. Since the services are still in the process of developing standards, 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 (implementation of the standards and conducting additional research to re-examine the standards) but that cannot be completed prior to the September 2015, FY2014 NDAA (sec 524) deadline for establishing gender-neutral and valid standards.[4] Instead, those elements will be likely next steps for the services once authorization is given to proceed with implementing the standards.

We therefore 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 DGCAR was lifted by the Secretary. But by using the four-step 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 in each step—placing less importance on whether they did them in precise order.

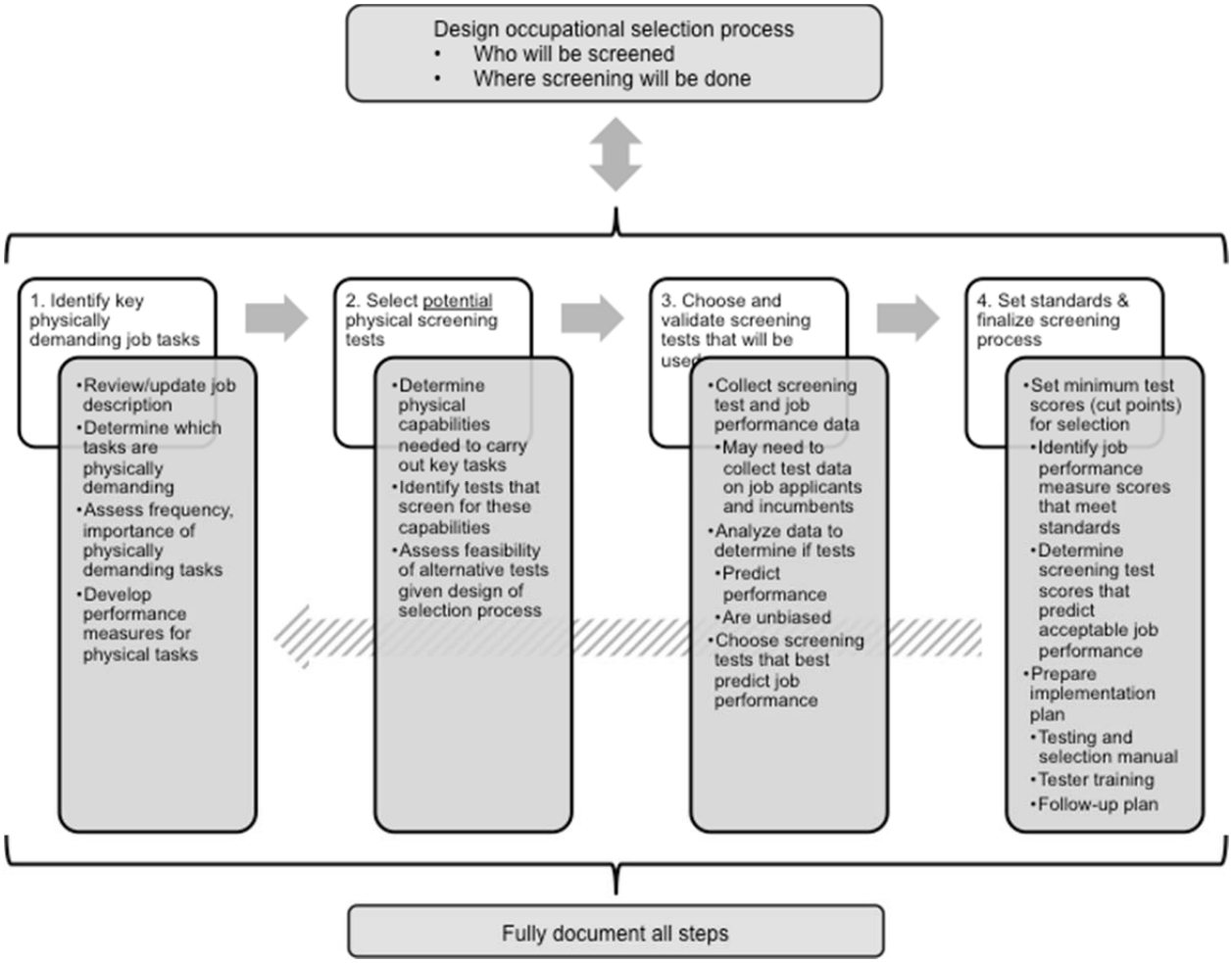
To understand the activities being undertaken, we met with representatives involved in the research in each of the services, reviewed documentation they provided summarizing the details of the work, and observed some of the data collection efforts. The representatives we sought out were those most knowledgeable 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. Much of the research documentation they provided to us is 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. Other

[4] This is an potential catch 22, where the standards need to be approved before they can be implemented and further tested to ensure they are working as intended (Steps 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 steps because approvals to proceed would be required before they could move on to Steps 5 and 6.

documentation included published and unpublished work conducted in prior years to support occupational standards established before DGCAR was rescinded.

Discussions took place over multiple 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 plans as the research unfolded. The summaries provided in the following chapters present the culmination of those discussions.

Figure 3.1. Physical Standards Development Process



As the work progressed over the multi-year 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 already 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.[5] Over the course of the project we observed changes being made in all of the service efforts as they progressed. As a result, only the final details about the work conducted to date or planned for the future are documented in this report.

The various discussions with each service took place by phone, VTC and/or in-person. The questions were unstructured, but we 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.

- Stage 1. Identify Physical Demands
  - How have you defined the physical demands of the job? Was there a job analysis? What was the process?
  - Who participated? Did you use subject matter experts? 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, etc.,) 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 under predicts 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?

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

  - 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? Have you conducted a study estimating the amount of improvement expected (e.g., as a result of basic training or technical training)?
- 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 February of 2015, none of the services had completed their work. None had arrived at the end of Stage 4. In addition, the Marines, Army and Navy had not yet established clear policy regarding what point in someone's tenure (prior to joining, upon joining, after boot camp, at the end of training, etc.,) the screening tests would be administered.[6] The timing of the administration of the testing will matter for setting minimum test standards in Step 4. Given that the Step 4 work is not complete, we cannot provide a detailed overview of the services process for establishing the final minimum tests cut points that service men and women must meet in order to 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 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 the final Chapter.

[6] 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 points in time starting as early as recruiting stations (for more on this, see the Air Force chapter).

Chapter 4. Army Combat Arms

,Four Army combat arms branches are currently closed to women in addition to the Special Forces occupation and Ranger assignments discussed in the following chapter. In total, this accounts 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 well over 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, they 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 following seven closed MOSs:

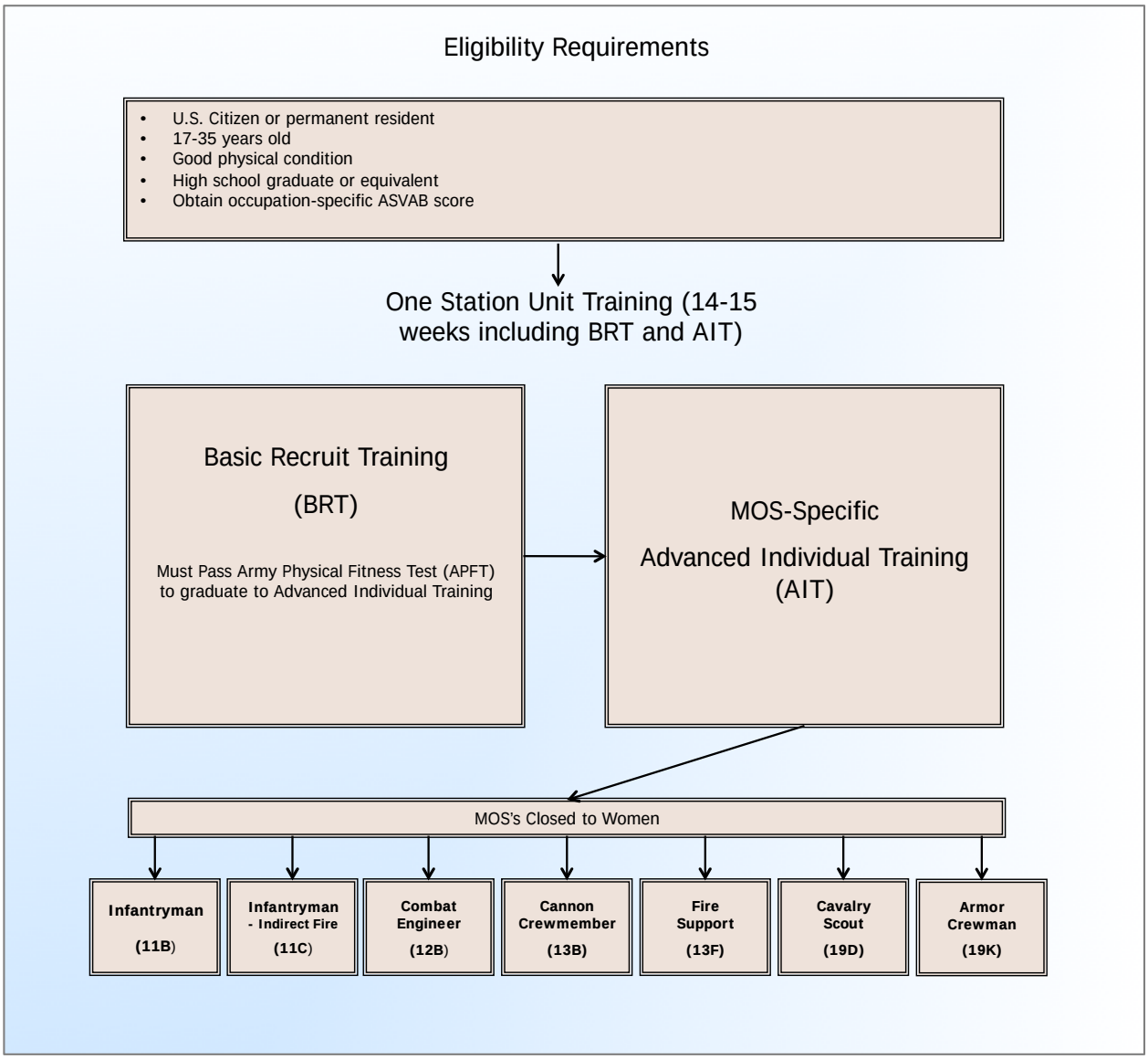
- 11B Infantryman
- 11C Infantryman-Indirect Fire
- 12B Combat Engineer
- 13B Cannon Crewmember
- 13F Fire Support
- 19D Cavalry Scout
- 19K Armor Crewman

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, at the time in which our data collection was completed. 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 plans 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: holding U.S. citizenship or permanent residency, falling between the ages of 17–35, being in good physical condition and moral standing, graduating from high school or holding an equivalent certification and scoring above a stated 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.

Figure 4.1. Eligibility and Training Requirements for Army Closed Occupations



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). The 10-week BCT includes a one-week Reception Week, followed by the Red Phase which focuses on building teamwork, the White phase which develops skills in marksmanship and rappelling, and the Blue Phase which continues weapons training and includes the culminating Night Infiltration Course and Rites of Passage. To graduate from BCT to AIT, recruits must pass the Army Physical Fitness Test (APFT). AIT takes place at one of 17 schools, depending on MOS, and includes both schoolhouse and field work components.

Table 4.1. Physical Tasks for Infantry (11B)

1. Frequently visually identifies vehicles and equipment at 1000 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 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 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 5-6 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 lbs pound person 3.5 feet as a member of a two Soldier team.

SOURCE: Army Pamphlet 611-21, Table 10-11B-1.

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 two-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 do not vary by gender for the sit-up event. For example, the optimum push-up score for 17–21 year old males is 71 and the optimum for 17–21 year old females is 42. If candidates do not pass the APFT towards the end of BCT, they will continue in a longer BCT cycle, with more opportunities 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. Each description includes a list of all physical tasks in the job and the frequency with which they occur. Table 4.1 shows the physical tasks for Infantry, which includes those in the 11B MOS.

Overview of the Army's Validation Effort to Date

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 develop a simplified set of simulations for inclusion in the criterion-related validation study. The third was a concurrent criterion-related validation data

collection using the simplified set of simulation activities. The first step is roughly aligned with our recommended first step (conducting a job analysis), whereas the second two steps most closely align with Step 3 (validating and selecting the selection tests).

USARIEM is conducting each effort separately for each MOS, beginning with combat engineers, which is the only MOS that had been completed and for which documentation was available at the time when our study's data collection period ended in March. The sections that follow therefore describe the process and preliminary findings for combat engineers. It is the Army's plan to follow the same steps for the other six MOSs currently closed to women; however, each effort will yield unique findings and therefore each should also be reviewed when complete.

The sections below describe the steps taken 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 Institutional Review Board (IRB) protocols for various elements of the research design and a preliminary draft of the technical report summarizing the results for the combat engineers.

Identifying Physically Demanding Tasks (Our Stage 1)

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. 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 tasks they identified 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, February 19, 2014):

- Conduct a tactical movement. Soldiers complete a 24-kilometer movement while wearing approximately 102 pounds of equipment (basic uniform, personal protective equipment, and 24-hour sustainment load).
- Move under direct fire (3–5 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 5 seconds 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 m away and assume prone position, rise, and sprint to fourth marker 15 m away and assume kneeling position, rise and, sprint to fifth marker 15 m away and assume crouched position, rise, and sprint to sixth marker 15 m away and run across the finish line.

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

Occupational Related Task	Military Occupational Specialty
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 25mm gun on the Bradley Fighting Vehicle	11B, 19D, 13F, 12B
Remove the Feeder Assembly of a 25mm gun on the Bradley Fighting Vehicle	11B, 19D, 13F, 12B
Load 25mm H-EIT Tracer Ammunition Cans onto the Bradley Fighting Vehicle	11B, 19D, 13F, 12B
Load TOW Missile Launcher on Bradley Fighting Vehicle	11B, 19D
Move Over, Through, or Around Obstacles	11B, 11C
Move Under Direct Fire (3-5 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 120mm Mortar	11C
Lift Emplace Cannon for 120mm 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 an Abrams Tank	19K
Load the 120mm Main Gun on an Abrams Tank	19K
Remove a Casualty from an Abrams Tank	19K
Transfer Ammunition with an M992 Carrier (M795 ME Rounds)	13B
Emplace 155mm Howitzer (Lift Wheel Assembly)	13B
Displace 155mm Howitzer (Lift Spade Trail Arm and Blade)	13B
Establish an Observation Point (Carry AN/PED-1(LLDR))	13F
Install/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	12B

SOURCE: Sharp, February 19, 2014

- 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 30-40 pounds of sand. Soldiers then move 26 sandbags (approximately 40 pounds each) 10 meters where they build a fighting position three sandbags in length and three sandbags in height. They have 26 minutes to complete the task.
- 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 Bradley Fighting Vehicle or Striker as quickly as possible.
- *Lift, carry, and install the barrel of a 25mm gun on the Bradley Fighting Vehicle (BFV).* As part of a two-man team and wearing a 83-pound fighting load, they lift and carry the 107-pound barrel of the M242 25mm gun for the Bradley Fighting Vehicle 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 as a team. Assistance from team members is permitted and 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 subject matter experts 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 description of the task, and to determine if 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 validation data collection effort (described later in this chapter).

Winnowing the 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 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 (females were recruited from across the Army). The females were similar to the male participants in average age (24 and 22 respectively) and military tenure (2 to 3 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., Bradley Fighting Vehicles were used in the *Remove a Casualty from a Wheeled Vehicle* activity). However, they 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 and self-reported run times for use in calculating VO_2max, heart rate, completion times, and distances obtained, as appropriate to the task. Participants were also given a questionnaire in which they 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 above. A set of four simulated tasks were 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 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 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 why each of the four new simulations was chosen 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 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 Bradley Fighting Vehicle 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 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 are as follows (Sharp, May 2014):

- *Casualty Evacuation from a Vehicle Turret*. Heavy lifting demonstrating muscular strength (Remove a Casualty from Vehicle, Carry and Emplace the Modular-Pack Mine System, Lift and Carry Rocking Roller During Construction of Baily Bridge, Load and Install Volcano). While wearing a 71-pound fighting load (full fighting load minus a

weapon), they squat, grasp the handles of the heavy bag level through a hole in a platform. The then stand and pull the bag through the hole and onto the platform. They first lift 50 pounds. If 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 25mm H-EIT Tracer Ammunition Cans on the Bradley Fighting Vehicle, 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 (to include 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 all relevant details such 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. Factors like heart rate, time to completion, and perceived exertion were measured as well. 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.

Identifying Potential Predictor Tests (Our Stage 2)

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 much of various types of physical abilities (e.g., muscular strength, muscular endurance, anaerobic power, trunk strength, etc.,) 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 US 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., 1 minute of push-ups, a 1 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 the following (Sharp, May 28, 2014):

- *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 the cones over, to the finish line, when the timer is stopped.

- *Standing Broad Jump*.

  Soldiers stand behind a line marked on the ground with feet slightly apart. A two foot take-off 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 backwards. 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 3-5 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, as fast as possible, 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.

Validating the Screening Tests (Our Stage 3)

Approximately 150 participants (male volunteers from the MOS and female volunteers from other MOSs) were recruited for the criterion validation—103 were male and 43 were female. Both groups were on average of similar ages (about 24 years old) and tenures (about 3 years).

During the study, participants completed the 4 simulations and all 12 predictor tests. Testing was completed in three different sessions with 24 hours or more between testing sessions. Researchers collected a wide range of data during the study such as heart rate, perceived exertion, number of repetitions, testing times, and other relevant test scores.

To determine the best predictor tests to include in the selection test battery, the researchers created regression models using the predictor 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 included only the best predictors among those that also meet 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 .60s to the high .80s.

USARIEM recommends that the Army select one of the three regression equations for use as the selection battery. They also recommend 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 under-prediction by gender). 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.

Our Assessment of the Army's Approach

The Army's approach appears to have some elements pertaining to our recommended Step 1, the job analysis. The job analysis 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 have not received copies of any written summary of the results of that work so we cannot judge whether the survey findings were consistent with the information provided by TRADOC. We also do not have detailed accounts of how TRADOC arrived at the final task list that they 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. We recommended that later versions of that technical report include such information.

The Army's effort did indirectly address our Step 2, in that they chose a wide variety of screening tools to include in the validation effort. Some are already used by other countries (as described in their preliminary technical report). Some were perhaps based on the results of subject matter expert 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. We recommend that final versions of their report explain this in greater detail.

The Army has also completed work that is in line with many of the recommended practices for our Step 3, validating the predictor tests. Their approach used a simulation-based criterion validation study in which the simulations were designed, measured and analyzed with care and attention to detail.

The work appears to stop at the completion of our Step 3. No plans have been described in the USARIEM protocols to address the establishment of score minimums for entry into the combat engineer occupation (our Step 4), and no plans to do so have been otherwise shared with us. As a result, we cannot comment on the methodology that the Army will use to establish the score minimums.

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, data USARIEM collected informed which of the original 12 task simulations could be combined to create 4 representative tasks for use in the criterion validation study. That links the outcomes being predicted 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 and the physical aptitude required to implement the task was low relative to the remaining 12 tasks. The rationales provided are 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 many 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. In addition, the testing establishes only the relationships at one point in time. It is not clear when the tests will be administered in the Army (the policy has not yet been decided); however, if the tests are administered as early as enlistment, the minimums established would need to be lowered to account for improvement resulting from training (such as Basic Combat Training) that would occur between the screening point and when soldiers are actually placed on the job. Lastly, the information contributing to the job analysis (which was used as the foundation of the study tasks) needs to be examined closely. If it was not established using a formalized and systematic job analysis process consistent with recommended practice, it should be redone to determine whether any changes to the job analysis information are needed. Given that TRADOC has (through their review of existing documentation) identified these as core tasks, it is likely that the tasks are in fact core to the occupation; however, a systematic process that similarly identifies the same tasks would be ideal.

Chapter 5. Army Special Operations Forces

Army Special Operations Forces (ARSOF) consist of Special Forces, Ranger, Special Operations Aviation, Psychological Operations, Civil Affairs, as well as Signal and Combat Service Support units. Many of the personnel assigned to these units are in occupations that are not specific to the ARSOF and some of these occupations are currently open to women. However, assignment to any of the ARSOF units is currently closed to women. The largest units are the Special Forces (SF) and Ranger units, so we discuss entry into these units in more detail here.

There is a dedicated Special Forces (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 Forces Command (USASOC) website (2015a):

> 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 WMD, and security force assistance. There are five active component Special Forces Groups 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 SF Group'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.

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. Here we describe selection and training to become a Ranger. The previous chapter focused on selection and training for closed occupations that are not assigned to Ranger or SF units.

Although there is no Ranger occupation, in general the process of becoming a Ranger resembles the process of entering the Special Forces occupation. Rangers first complete training in a range of occupations instead of all completing infantryman training. USASOC (2015b) describes the 75th Ranger Regiment on their 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.

Like the Special Forces, they also work in 12-person ODA teams, with each member of the team contributing a different occupational area of expertise.

Special Forces units account for around 7,000 closed positions, and the Rangers account for over 2,000.

USASOC has accepted primary responsibility for establishing gender-neutral standards for both of the Army's special operations MOSs in response to lifting of DGCAR, and those efforts are described later in this chapter.

Occupational Assignment and Screening in USASOC

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 join the Rangers and Special Forces applicants must:

- Achieve passing scores on the Army Physical Fitness Test (APFT), which includes a 2-mile run, pushups 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 12-14 minutes and completing 80 to 100 sit-ups and push-ups
- Have no physical limitations
- Be a male aged 20-30 (for Special Forces)
- Be a U.S. citizen with a high school diploma
- Obtain a General Technical score of 110 or higher on the Armed Services Vocational (ASVAB) Battery for the SF MOS and 105 or higher to join the Rangers; for the SF, also have a combat operation score of 100
- Qualify for secret clearance
- Qualify and volunteer for Airborne training
- For SFs, have 20/20 or corrected 20/20 vision.

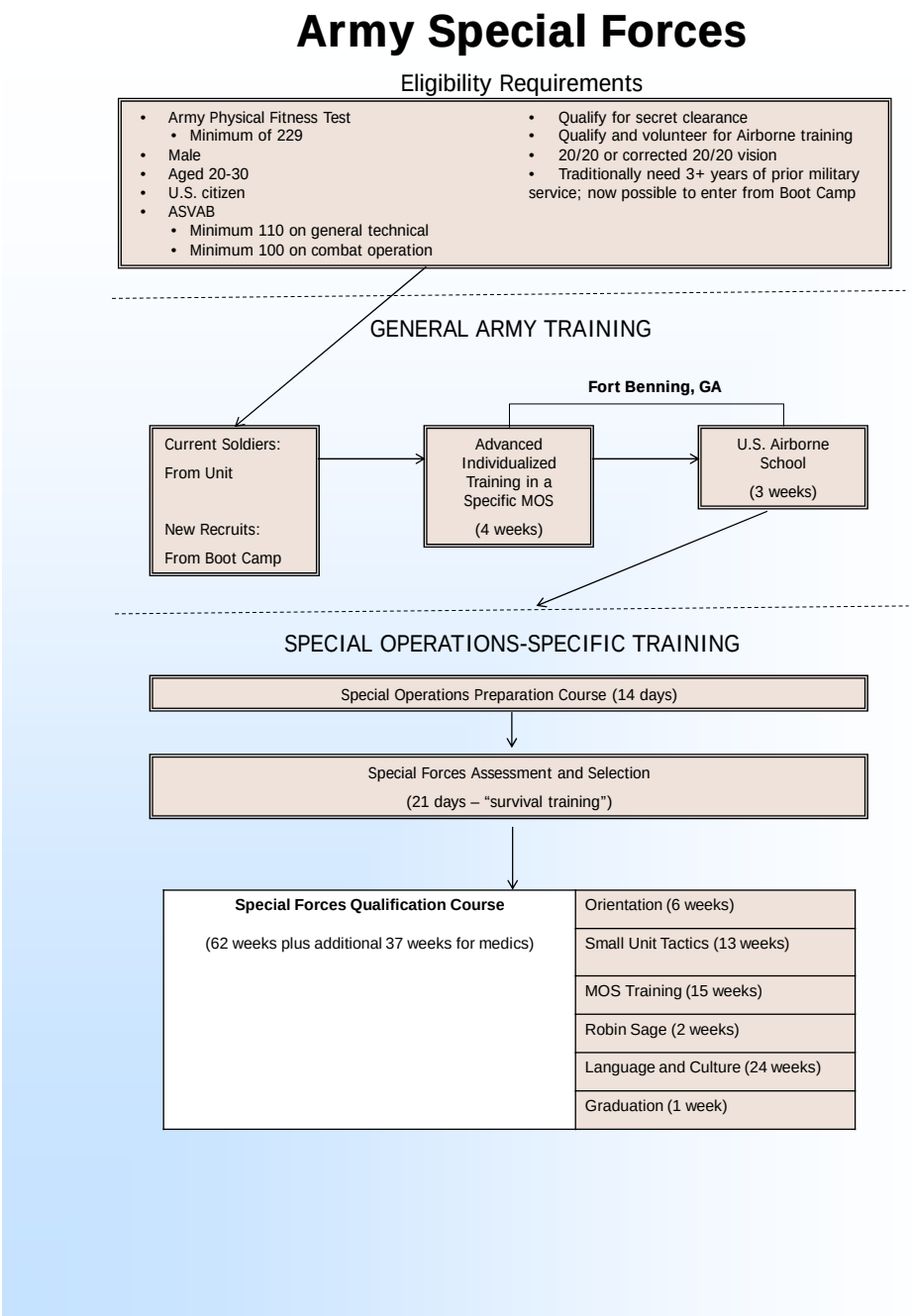
These are the minimum requirements to submit an application, not to qualify to enter SF training. For most occupations, soldiers are selected for an occupation if they meet the stated requirements and there is a training slot available in their time frame. In contrast, selection into the Army special 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 (to include physical fitness test scores). Applicants are accepted into the assessment and screening process and subsequently the Special Forces Qualification 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 above), that pre-screening 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 trainees where a large part of the selection process takes place.

For SFs, the first stage of training is 16 weeks of infantry One Station Unit Training, which includes both the Basic Combat Training course and Advanced Individualized Training course in a specific MOS (Table 5.1). Upon completing One Station Unit Training, special forces trainees enter U.S. Airborne School/Basic Airborne Training (3 weeks, focused on parachuting and landing safely) at Fort Benning, GA. Following that is the first Special Forces-specific training course, the Special Operations Preparation Course (SOPC), which is a 19-day course held at Fort Bragg that focuses on developing physical fitness and land navigation skills. After SOPC, recruits begin the 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, pushups, obstacle course, marches, land navigation/orienteering and leadership and teamwork.

Figure 5.1 Army SOF Screening Lifecycle



For those Special Forces trainees who make it through SFAS, the remainder of training is called the Special Forces Qualification Course. The course is divided into 6 phases over 62-weeks (with an additional 37 weeks for medics), which includes:

1) Orientation (6 weeks): orientation to the full course plus training in unconventional warfare, special forces history, organization, task/core activities, capabilities, methods of instruction, patrol orders and troop-leading procedures

2) Small Unit Tactics (13 weeks): advanced marksmanship, counterinsurgency, urban operations, live fire maneuvers, sensitive site exploration as well as Survival, Evasion, Resistance and Escape exercises
3) MOS Training (15 weeks): Special Forces-specific training in the MOS including language training, Special Forces tasks, Advanced Special Operations Techniques and intra-agency operation
4) Collective Training/ Robin Sage: During this 2-week 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): 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, Urdu
6) Graduation (1-week) out-processing, soldiers now wear "Green Berets" as Special Forces Soldiers

The initial training pipeline for Rangers is shorter than that of the SFs (shown in Figure 5.1). Like the Special Forces they first complete Basic Combat Training and Advanced Individualized Training 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 8 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 that occurs during Ranger training. Those that 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 continued assignment to 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.

Typically only about 45 percent of recruits complete the entire Special Forces or Ranger training pipeline, with a large portion of the losses of trainees happening during SFAS and

RASP.[7] 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 training in these MOSs, beginning in 2015 USASOC opened a number of seats across the January through April Ranger training cohorts to female volunteers. Two women have successfully completed RASP in August 2015, but they cannot be assigned to the Ranger Regiment until it is opened to women.

Army Process for Establishing Standards

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.

USASOC has, however, initiated a new effort to validate SFAS and RASP in direct response to lifting of DGCAR. To accomplish this, USASOC turned to the Office of Personnel Management (OPM) and the Naval Health Research Center for assistance. The overarching goal for the work (as stated by USASOC) is to establish the relationship between the training tasks required in the courses and those documented in DA PAM 611-21 (Personnel Selection and Classification Military Occupational Classification and Structure Pamphlet, 2007).

Job Analysis (Our Stage 1)

The last job analysis to be completed on 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 Special Forces and Ranger occupations. The goal of the research effort is to ultimately determine the knowledge, skills, abilities (KSAs) required of Special Forces and Ranger soldiers. OPM's job analysis approach includes reviewing background occupational information, as well as conducting site visits and administering a survey.

The timing of the OPM effort was such that we were not able to acquire details on the methodologies they are using for their analysis or documentation of the results. The anticipated results have been described by USASOC as including a list of the tasks, competencies, and physical abilities required on the job. They also anticipate that OPM will provide detailed documentation on the method and results once their work is completed.

---

[7] Only 42 percent of those who attempted Ranger School between 2010 and 2014 completed the course, with the majority of the failures (62 percent) occurring due to the Ranger Physical Assessment (Airborne and Ranger Training Brigade website, 2015). 36 percent of students fail in the first four days.

Establishing the Link between the Selection Criteria and Job Competencies (Our Stage 3)

According to OPM's scope of work, their final deliverable, due by the end of the third quarter of FY2015, 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 would 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 not unfairly discriminatory. Although USASOC has noted that OPM will use statistical techniques to accomplish this, details on those techniques and the data underlying them have not been provided. We infer 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 they would be providing data to the Naval Health Research Center (NHRC) Exercise Physiologists who would assist in a criterion validation and standards validation process (time and data permitting). The Army representatives directed us to speak with NHRC regarding those analyses. However, because no data had been provided to NHRC when we interviewed them, NHRC could not yet describe exactly what data would be included or how that data would be used. As a result, our work ended before we were able to determine how NHRC would be contributing to the effort, if at all.

Lastly, although candidates are pre-screened on physical fitness scores (along with other information) prior to even being allowed to enter training for these occupations, the Army had not considered that to be an important element to examine in their work in response to the lifting of DGCAR. As a result, at the time in which we completed our interviews, their work does not include any efforts to validate that part of the screening process.

Our Evaluation of Both Stages

The 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 before our project work ended. Therefore, the validation effort described above lacks the details necessary to evaluate the reasoning, logic and methodological soundness of the approach.

From the information provided, it appears that the Army is relying heavily 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),

etc.,—we cannot comment on the soundness of the job analysis findings. We also do not have enough information yet to determine how the job analysis results will be used to inform which tests are most appropriate for use as screening criteria before and during training, nor can we determine how minimums on those tests will be established or how bias on the test will be examined.

Although few details on the methods OPM is applying in this context are currently available, OPM has a long history of work in the areas of job analysis and validation of selection practices. They also serve as a resource to the public to help in designing valid and unbiased selection practices. For example, on their website (OPM website, 2015a) they state 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.

And 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 also provides a detailed account of what they describe as their job analysis methodology in the 2007 Delegated Examining Operations Handbook as a reference for the public on methods for job analysis. That methodology includes a highly detailed and structured questionnaire administered to subject matter experts (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. Additionally, it is still not clear how such job analysis information alone could be used to determine which tests are most valid, what the minimums on the tests should be, and whether the tests are biased against any relevant groups.

A well-engineered and executed job analysis lays the foundation for amassing evidence to support a selection system (our recommended Step 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 Step 3). That link could be established using content validity information if the content validity evidence is strong. However, we cannot know how strong that evidence would be until we see the details of the method used for establishing it, as well as the results of any data collection efforts and the tests selected for screening applicants.

One important detail to establish those links, for example, concerns confirming that the training itself can be used as a fair and accurate screening tool. We do not know how consistently the standards are applied from one person to the next or one training class to the next. It is possible that training in one class is harder than training in the next (e.g., due to 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 at determining who and who will not be competent at meeting 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 could be used for such screening. Evidence to support this would need to be part of the content validation process, both to inform Step 3 of our recommended process (validating the selection criteria) and Step 4 (establishing minimum standards).

The OPM job analysis as it was described to us does 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 Step 2 is being addressed by the Army's current approach. It is possible that activities that are 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. We do not know to what extent OPM plans to address this.

Additionally, like the other services, no clear information regarding methods for establishing the minimums on the screening criteria has been provided (Step 4). However, we expect that OPM will provide detailed documentation of their process once it is complete, which will help OSD in better understanding the details of the work.

Lastly, little attention has been paid to ensuring that the screening process used to determine who is selected for training is also validated. The OPM job analysis findings can likely help inform decisions about the pre-training screening process as well as the training screening processes. We recommend that the Army explore this further. OPM's past experience with job analysis and the use of job analysis information leads us to suspect that their approach to the job analysis will be defensible, by reputation alone. Nevertheless, the details matter, and at this point in time none of the details of their method have been disclosed.

Chapter 6. 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 non-primary 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, falling between the ages of 17–29, being in 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.

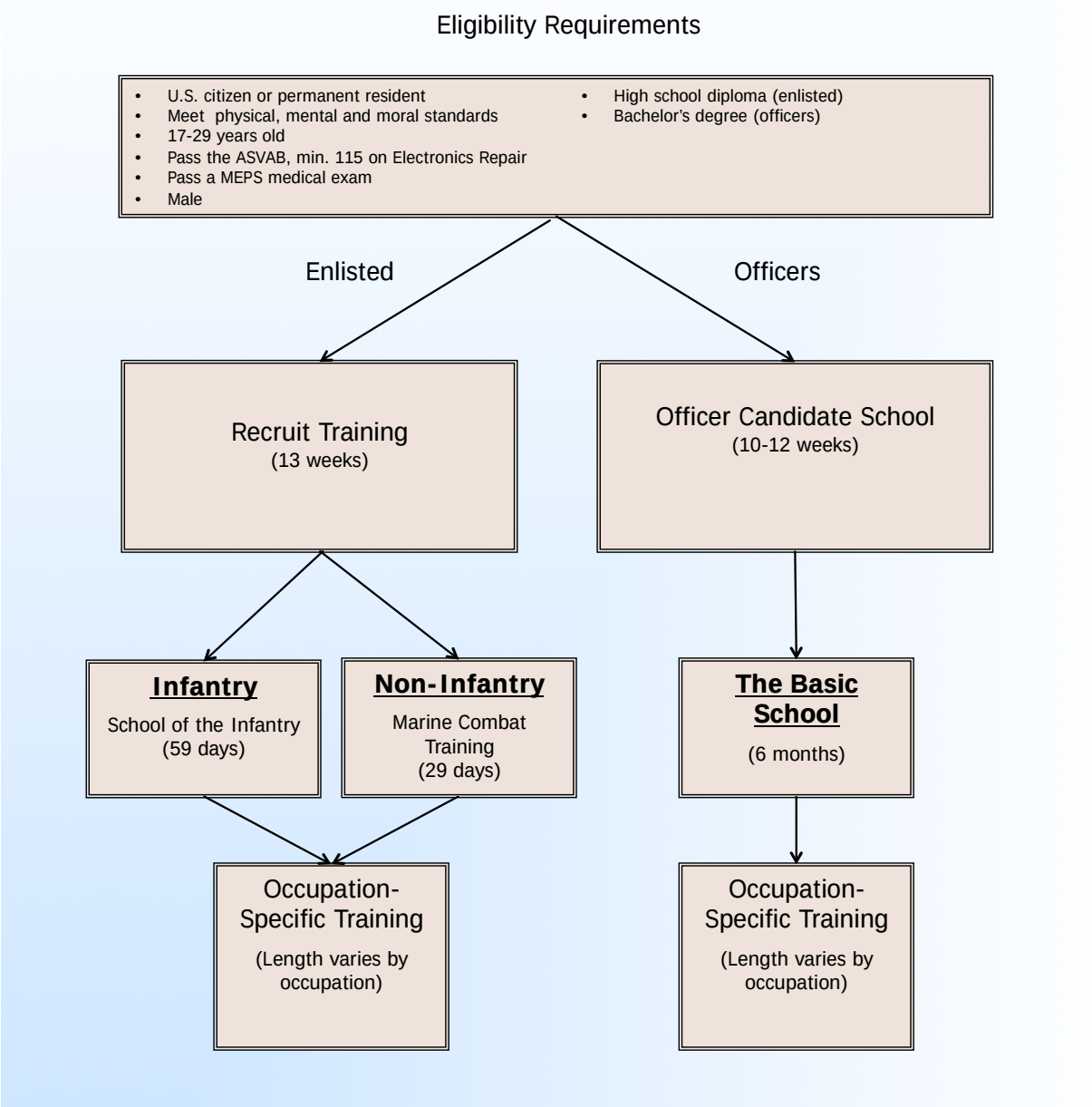
Interested males enter the closed enlisted occupations through their local MEPS. They contract for an occupational career field when they join and receive their specific occupation assignment during recruit training. Recruits choose from the career fields taking new recruits at the time they will enter and for which they qualify based on their aptitude test scores. No occupation-specific physical qualifications current exist for entry-level occupations. However, to demonstrate their "good physical condition," before enrolling in Recruit Training all enlisted recruits must pass an Initial Strength Test by completing 2 pull-ups (males) or a 12-second flexed arm hang (females), 44 sit-ups in two minutes, and a 1.5-mile run in 13:30 minutes (males) or 15 minutes (females). However, Marines recruiters recommend to recruits that they strive for well beyond these minimum requirements. No other physical screening is conducted currently 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 non-infantry 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 10–12 week Officer Candidates 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 time 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 half way through The Basic School.

Figure 6.1 Entry and Training Path for Marines Combat Arms Branches



Given the current 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 pre-accessions (recruiter-administered) screening to physical screening conducted during recruit training and prior to MOS school assignment" (Secretary of the Navy, 2013). 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.

Overview of the Marine Corps' Validation Efforts to Date

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

7. **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 candidate physical screening tests with performance in proxy tasks of the most physically demanding tasks identified in the first step. This effort was completed by Training and Education Command (TECOM) at the end of calendar year 2013.
8. **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. TECOM is surveying participants to assess their reasons for volunteering, attitudes prior to training, and experiences and attitudes if they drop out or when they 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.
9. **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 they will fill in 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 will be evaluated as they rotate through a series of simulated events. The evaluation will compare the ability of units with no women

versus one or two women to meet performance standards in the events. The performance of individual participants will also be measured, compared across units with differing gender mixes, and correlated with individual physical characteristics. Other outcomes to be evaluated include attrition and injury rates; medical readiness and deployability; and cohesion, morale, and discipline. The results will be used along with the results of the first effort to finalize screening procedures and criteria for closed occupations and for assigning marines in open occupations to ground combat units. At the time in which data collection for this project ended, the integrated task force had not yet started their data collection efforts.

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

Methods Applied in the First Criterion-Validation Study

The Marine Corps assigned responsibility for carrying out this study to its TECOM. TECOM completed this effort in December of 2013.

Identifying Physical Demands (Our Step 1)

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. The Marine Corps established the T&R manuals in 1995 and in 2011 developed 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 (United States 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 lists 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[8] 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 non-deployed environment, and whether there are any equipment changes that affect the tasks.

Additional data is 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 other tasks.

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

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 required to qualify for the occupation. However, in discussions with both 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 explore the issues further.

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 subject matter experts 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. Subject matter experts 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, 10 entry-level enlisted occupations were included in the study. Table 6.1 lists the 10 occupations by occupational field. 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	Military Occupation Specialty (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) 1812 M1A1 Tank Crewman 1833 Assault Amphibious Vehicle (AAV) Crewman

Using the Marine Corps' Existing Fitness Tests (Our Step 2)

TECOM opted to use the existing Marine Corps fitness tests (the PFT and CFT) as the predictor tests in their first data collection effort. The PFT has been the Marine Corps' general fitness test for over 30 years. Responding to a high rate of non-combat injuries in Iraq, the CFT was developed in 2008 to supplement the PFT with a test that is more combat related and to 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. The components of the two tests are:

- 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 Can 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, they used inputs from subject matter experts 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 2 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, 18–23 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 over 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.

Correlating PFT and CFT Scores with Physical Task Performance (Our Step 3)

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 tasks shown in Table 6.1 according to the type of physical capability required. They 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, they opted to develop a single set of screening criteria for all of the closed occupations. They 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 their own occupation. Given that the proxy tasks were designed to reflect only the most physically demanding tasks across all of the occupations in Table 6.2, the final set of proxy tasks may not reflect the demands in any single occupation in that table.

Table 6.2. Proxy Tasks for Physically Demanding Tasks in Marine 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 lb. Assist pushing crewman out of turret from below – 115 lb.	Clean & Press	Lift bar with weights to shoulders and then lift above head • 1 repetition each to maximum completion at 70, 80, 95, 115 lb.; participants could elect to skip lower weights • 6 repetitions at 65 lb. in 1 min.
Lift heavy object to lower height	19	Replace track block on M1A1 tank – 60 lb. Lift light armored vehicle strut assembly – 3-man team at 135 lb. each	Dead Lift	Lift bar with weights to knuckle height • 1 repetition each to maximum completion at 60, 70, 80, 95, 115, 135 lb.; participants could elect to skip lower weights
Lift and carry heavy object	2	Lift & carry 155mm round 50 meters in 2 minutes Lift & carry 100 lb. general mechanics toolbox	155mm Lift/Load	Lift & carry 155mm replica artillery round weighing 95 lb. 50 meters, wearing fighting load • 1 repetition in under 2 min.
Lift and load heavy object	1	Load M1A1 rounds (gunnery skills test, requires 5 rounds in under 35 sec.)	120mm Lift/Load	Lift 120mm (55 lb. each) replica tank rounds off 20” box, flip, and stack on 2nd 20” box • 5 repetitions under 35 sec.
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” assist box, wearing fighting load • 1 repetition

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—120mm 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 current body armor, helmet, gun, and ammunition). Each of the five proxy tasks was performed in a single or limited number of repetitions, and therefore do 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 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 about 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 above 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 they 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 percent 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 men and 99 percent of women completed it. All men were also able to complete three of the four other tests, and roughly 70-80 percent of female subjects did so. 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. Of the nine physical job tasks for which the clean and press serves as a proxy, three require lifting a weight of 100–115 pounds, two involve weights of 60–85 pounds, and four involve under 60 pounds. Just about all men tested could handle weights up to 80 pounds and over 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.

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 depends 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/CFT. 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 three-mile run, and 0.37 for crunches.

The NHRC/TECOM report concludes that the CFT events are overall 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 and the run predicts better than crunches. The report concludes that 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 their bivariate correlations, the researchers recommend that applicants for combat arms occupations be screened using an enhanced Initial Strength Test (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 Screening Tests and Minimum Scores for Selection (Our Step 4)

The NHRC/TECOM report recommends that a version of the current initial strength test be used to screen Marine recruits interested in the infantry field. 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 cut-off scores for allowing recruits to enter the infantry field based on the scores for the three screening events that were posted by test subjects able to perform the proxy tasks. Specifically, the cut-off score for each screening event was set at the mean score for the lowest decile of Marines able to perform the proxy tests. Cut-scores were calculated based on the data for test subjects who successfully completed at least four of the five proxy tasks (75 percent on the composite performance score) and for the smaller group of test subjects who successfully performed all proxy tasks (100 percent).

Using the 75-percent proxy task performance standard, the calculation yielded a minimum of only a single pull-up but the report recommended using a higher minimum of three pull-ups, currently required for male Marines to pass the PFT. Otherwise, the cut-scores represent a more stringent standard than the currently required to pass the PFT and CFT. Using the 100-percent proxy task performance the cut-scores reflect a substantial increase in the minimum standard compared to the PFT/CFT, for male and female Marines. One-third of the subpopulation of test subjects meeting the 75-percent performance standard was female and 8 percent of the subpopulation meeting the 100-percent performance standard was female. Among male subjects, all would qualify using the 75-percent standard and four-fifths would qualify using the 100-percent standard; among women, the qualification rates would be 56 percent and 7 percent,

respectively. Not surprisingly, where the cut-scores are set for this suggested screening test would have a greater impact on women than on men.

Our Evaluation of the Approach

The procedures outlined in the Terms of Reference for the Marine Corps Ground Training and Readiness Group (TRMG) generally correspond to the procedures associated with our recommended first step, 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 recommended Step 2, the Marines did not fully address it. The decision to rely on PFT/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 they cite supporting this decision evaluated how well the six PFT/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 Step 3 (establishing a relationship between the tests and on-the-job performance), 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 used to analyze the data. We discuss each of these in turn.

First, the study used a simulation-based criterion 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—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 for the tasks. These tests therefore 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 and therefore 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 task 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 will have to be taken into account in deriving occupational qualification standards from the task performance results. TECOM recognizes that additional data collection and analysis will 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 on individual Marines will provide a more definitive validation of the standards than is possible in the standards development process.

It is also worth noting that the simulations were intended to apply to all closed occupations regardless of whether that occupation required the task it was intended to simulate, 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. For example, 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/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/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 matters. The TECOM study collected concurrent validity evidence, meaning that the data on the predictor test (the PFT/CFT) and job performance measures (proxy task results) were collected at about the same time for the same individuals (the PFT/CFT scores were the most recent scores taken 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 initial training, when occupational assignments are made. However, for screening at the recruiting station (or Military Entrance Processing Station), the minimum scores should be set lower to allow for improvement in physical capability during

basic training. If the screening is also done to re-qualify 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, because 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 women are allowed into infantry and other ground combat occupations and required to do pull-ups, their ability to carry out the job tasks proxied by the clean and press test should improve.

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. It is likely that most who chose this option were male. In theory, this could affect their performance in lifting heavier weights relative to what it would have been had they lifted all possible weights, due to fatigue effects. In practice, because a high fraction of men could perform all lifting tasks, the option may have had little effect on the results.

Third, the way in which the validation study data are analyzed matters. The validity analysis relied on correlations between scores on the individual PFT/CFT tests and either each proxy task result (pass or fail) or the combined result on all proxy tasks (percent 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/CFT scores, it is inevitable that analysis of the pooled male and female data will show a positive correlation between the PFT/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 is unfortunate as, gender is likely at least as highly correlated with task performance in these data as the PFT/CFT test results are. Thus, we cannot rule out the possibility that the correlations they 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, and for a test to be considered valid and fair for both groups, those relationships should hold within gender. 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 gap in the study results, one that raises questions about whether the screening minimums would even be considered valid for the male population.

An analysis of the data for women only would provide information on the validity of the PFT/CFT for determining which women may be capable of performing physically demanding tasks in currently closed occupations. However, this would not support the validity of the PFT/CFT for setting gender-neutral physical standards. As we discuss in our overview of methods for developing occupational physical standards (Hardison, Hosek, and Bird, 2013), gender neutrality means employing the same tests and selection criteria for men and women *and* 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 as well as women. Such testing should include male subjects from the same occupations that the female subjects come from and include physical task performance measures that can distinguish different physical capabilities among men.

The NHRC/TECOM report concludes that, with a few exceptions, the CFT/PFT 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 .30 to .40—is not applicable in this context. The source they referenced for the standard was a RAND report (Hardison, Sims, and Wong, 2010) which 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 .30 to .40. In contrast, tests of physical aptitudes when correlated with physical simulation activities in a laboratory setting can show correlations ranging from .60's to .90's, depending on the complexity of the simulations, the timing of the predictor and outcome measures and many other factors. Holding all of these other factors 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). Under these circumstances, the bivariate correlation of each IST+ test and the proxy performance tests to a considerable extent captures the effect of the scores on the other two IST+ tests. Multivariate (regression) analysis incorporating all PFT/CFT scores would normally be employed to estimate the contribution of each screening test to predicting task performance. Evidence for the potential importance of multivariate analysis is contained in a study of lifting and carrying task performance by naval personnel, which found that bivariate correlations measured between dynamic strength or

anaerobic power tests and performance on physically demanding tasks largely disappeared in analysis controlling for the effects of static strength and aerobic capacity (Vickers, Hodgdon, and Beckett, 2008). However, multivariate analysis may not be feasible here because of the high correlations between the IST+ components and the limited variability in the task performance data for males.

The researchers conducted regressions to identify the best combination of predictors and to examine the gain in prediction over more easily implemented tests. Although running regressions for this purpose is generally consistent with recommended practice, it is not clear whether the recommended tests would be significantly different if the regressions had been run within gender. This is also a critical gap in the analysis that should be addressed.

With respect to our recommended Step 4 (establishing minimums), TECOM's approach was simple to understand, although because of some of the limitations described above, they 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 cut-scores calculated in the NHRC/TECOM report essentially 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 cut-scores will be set too low or too high. Too-low scores will qualify individuals who fail to pass training or who perform poorly on the job. Monitoring training and job performance over time should quickly identify any need to raise the cut scores. Too-high scores will be more difficult to detect.

The NHRC/TECOM report does not present data on the accuracy with which the suggested screening test identifies who can perform job tasks and who cannot, given the calculated cut-scores. Any screening test will yield 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 will yield fewer false positives and more false negatives. Those qualified for the occupation will 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 percent 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. If the Marine Corps adopts their recommendation, they would have to put forth evidence showing that Marines in combat occupations are regularly expected to perform not only the physically demanding tasks associated with their own occupation, but also that of the other ground combat occupations.

Methods Applied in the Ground Combat Element Integrated Task Force (GCEITF) Study

The research for developing screening tests and standards for closed combat occupations, described in the previous section, is complete, but a final decision on a physical screening test will not be made until information from the Ground Combat Element Integrated Task Force (GCEITF) is available toward the end of FY2015. The GCEITF study was in progress when we completed data collection for this report, so we do not know what the results will be or how they will inform decisions about the final screening process for selection into combat occupations. We do, however, have detailed information about the design of the experiment and the analysis plan, so we can evaluate how the results will relate to setting standards for selection into the Marine Corps' closed occupations.

The entire study is 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 below relates primarily to our Step 3 (i.e., validation of the predictor tests).

Study Objectives

Detailed objectives of the GCEITF are stated in the research protocol for the study prepared by the Marine Corps Operational Test and Evaluation Activity (MCOTEA), which is overseeing the experiment with assistance from researchers from the Center for Naval Analysis, the Naval Health Research Center, and academic consultants.

This study had two broad objectives and its design was not targeted to establishing physical standards. The first objective explores whether units assigned some women perform any differently than units that are all male. Note that all participants have completed occupational training so this objective is not about which men or women can complete training and perform in a unit. The second objective focuses on individual abilities and both individual and unit performance, so the evidence collected toward this objective may be relevant for establishing and validating gender-neutral standards. The second objective is the 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 design is intended to explore the relationships shown in Figure 6.2—individual physical characteristics impact mission outcomes through individual readiness, proficiency, and conduct. Thus, 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. Their premise is that someone struggling to carry out assigned tasks and affected by extreme fatigue may perform less well in these other dimensions than that person otherwise would.

Figure 6.2. Hypothesized Relationship Between Individual and Unit Attributes in the Ground Combat Element Integrated Task Force Study

Sample

The experiment is 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, assault amphibious vehicles). Each type of unit conducts multiple trials of a series of mission events to determine whether and how outcomes differ for integrated versus male-only units. Each trial involves three units of the same type performing a set list of tasks associated with a particular type of mission. The matched units include a single male-only unit, a unit with a single female member, and a unit with two female members. In the integrated units, the ratio of men to women ranges from 11:1 to 1:1 depending on unit size and whether the unit has one or two women assigned.

A random sample of volunteer participants is selected for the units and for specific roles within units for each trial. Sampling is by replacement, so an individual Marine has the same chance of being selected for each trial, unit type, and role. The only exception is when more than one trial is conducted at the same time. This procedure is designed to ensure that unit leadership and the individuals assigned to the units do not affect the overall comparison of performance by unit type (male, one female, two females). The trials are conducted at the same times and under the same conditions for the different unit types.

The number of trials per mission event and the number participants was based on a power calculation using information derived from operational tests if available and from live demonstrations of the events otherwise. The power calculations were designed to detect a 30 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 non-availability, the actual sample is 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 under nine years of service
- Full-duty status
- Closed MOS volunteers (men only): capable of achieving the third class Physical Fitness Test requirements for males age 1726 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 women volunteers. The women selected first 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 is on-the-job experience. Male participants likely include some Marines just out of training, but many will have had unit experience and some will have been deployed. Since women cannot currently be assigned to regular ground combat units, 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. In contrast, the 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/CFT elements, height and weight) and perhaps other individual characteristics such as ASVAB scores.

The study protocol notes that, because participation is voluntary, participants are self-selected. If female participants volunteer based on their interest in serving in ground combat occupations, the study population would likely be representative of women who would seek to enter these occupations when they are opened. Male participants ideally would be representative of the population of Marines serving in ground combat units, but it is unclear whether the volunteers for this study are likely to represent a cross-section of this population. For example, as noted in the protocol, they may disproportionately represent Marines who either support or not support opening ground combat occupations to women. The researchers will have considerably opportunity to closely observe the study participants and may be able to determine what motivated participation.

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 worked with a number of Marine Corps organizations: Plans, Policy, and Organization; TECOM; Ground Combat Advisory Groups; and the 1st, 2nd, and 3rd Marine Divisions.

Table 6.3. Mission Events by Occupation for Ground Combat Element Integrated Task Force Study

Occupation (MOS)	Events
Machine Gunner (0331)	• Provide suppressive fires with medium machine gun • Provide suppressive fires with heavy machine guy
Rifleman (0311)	• Conduct ground attack (with MOS 1371) • Conduct defensive operations • Conduct dismounted movement in mountainous terrain
Mortarman (0341)	• Provide indirect fires with 60mm • Provide indirect fires 81mm 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 & 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
Assault Amphibious Vehicle (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 & remedial actions on weapon • Conduct simulated reload • Evacuate wounded crewman (2 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

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 encompass 48 hours with an attack on the first day, a seven kilometer march and holding a defensive position on the second day, nighttime bivouacs both nights, and various non-experimental 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 is carrying 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 varies from 14 to 40, as required to achieve the desired statistical power. A male-only unit, a unit with one female, and a unit with two females carry out each trial. As we described previously, the individuals in each of these units are randomly chosen from the pool of participants for that occupation(s) by gender for each trial.

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

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) • Self-reported maximum workload (7-item scale)
Measures spanning multiple events	• Individual readiness: % days available for duty • Unit readiness: % 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

The University of Pittsburgh's Neuromuscular Research Laboratory is collecting extensive physiological data on GCEITF participants, including measures of flexibility, aerobic capacity, and stress and they will administer the PFT/CFT and other screening tests. We were not given a description of the physiological data collection, including important details such as when the data were to be collected (e.g., before and after collective training, during events or between events). Scores on earlier PFT/CFT tests from basic training and occupational training are also available

in personnel records. These data will 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.

Analysis of Experimental Data

Earlier, we described the two principle 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 are more fully developed for that purpose.

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

1. 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 percent of each A1M1 tank crewmember) measure either elapsed time or rate of movement performing a task, or less frequently on-target firing percentage. The other individual measures are collected at the event level instead of the task level and consist of self-reported fatigue and workload level during the event.
2. Similarly, comparison of the distributions of individual performance and conduct during the event overall, as evaluated by unit commanders during the events. The protocol does not indicate whether the distributions will be done for all occupations and events combined or by occupation, event, or both.
3. 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).
4. ANOVA analysis to determine whether individual task proficiency (the outcomes analyzed in (2) above) is correlated with collective performance in events. As with the regression analysis, the ANOVA analysis will control for other variables.

Overall, the proposed analytic methods are reasonable given the experimental design and, in the context, the objectives related to the measures of individual and unit performance during the experiment. The data record multiple individual performance measures for the same individuals in different trials of the same event and across different events. The individual participants are randomly combined with other participants to form units and randomly assigned to different

roles in the units. The protocol anticipates that randomization will assure that 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 are unlikely to be identical units (the same members assigned to the same roles).

However, there is still reason to question whether the observations will meet be 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 will be correlated and the statistical power for the experiment will be lower than expected. Similarly, persistently higher or lower performance by the same individual will cause some correlation across unit-level performance measures for all units to which the individual is assigned. The protocol discusses 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

The analyses proposed in the MCOTEA protocol do not describe a clear plan to employ any of the selection test validation methods in our recommended step (see Figure 3.1). That is, they do 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 does mention that analyses of the relationship between individual physical characteristics and performance outcomes will be conducted, but it does not describe the specific methods that will be used, nor does it state how the results would be used.

However, MCOTEA will have the data necessary for conducting this type of analysis. Regression analyses (similar to those they propose in items (3) and (4) above) could be conducted to identify the best predictors or a combination of predictors for use in selection. The potential predictors could include PFT/CFT scores (taken at various points in time), the initial strength test (a subset of the PFT/CFT), and other physical aptitude measures collected during or prior to the study. Outcomes could include any of the 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 initial strength test scores in basic combat training 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 they propose in items (3) and (4) above, other factors, such as weather, could also be controlled for in the regressions. These regressions should also be explored by gender to determine whether the same physical test scores predict equally well for men and for women.

Lastly, care should be taken in interpreting findings showing that women do not perform as well in the unit events as men. In theory, all of the male and female participants will have

graduated from the same (but segregated) MOS-specific training. If there are performance differences in the unit events during the experiment, it could raise questions about whether the gender-segregated training was actually comparable in difficulty. 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 large proportion of participants (male or female) fail to perform satisfactorily, it should raise questions about why the training did not adequately prepare them to do the job. If so, assuming that the expectations for what constitutes satisfactory performance in the unit events are 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.

Conclusion

The Marine Corps will rely on the results of two major studies in developing physical standards for its closed ground combat occupations. The first study explored the correlation between scores on the PFT/CFT tests and simulated individual physical task performance. This study generally followed the basic steps 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 recommended steps, we identified several limitations in the data and the analyses that could affect the validity of the suggested standards.

The CGEITF could provide additional data and analysis that may 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 their 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 cannot fully evaluate how useful the experiment will be.

Chapter 7. Marine Corps Special Forces

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 (SOCSs) and Special Operations Combat Service Specialists (SOCS-Ss) account for the remaining closed positions. They are instead temporary assignments or billets filled by members of the broader Marine Corps community.

Occupational Assignment and Screening

The CSO/SOO and SOCS training and selection pipelines are displayed in Figure 6.1. 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.

SOCS personnel are first assigned to MARSOC through award of the 8071 secondary MOS 3-12 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: explosive ordinance disposal—6 weeks, communications—12 weeks, intelligence—10 weeks, joint terminal attach controller—4 weeks, multi-purpose canine—10 weeks, special amphibious reconnaissance and independent duty corpsman—13 months. Those who complete the training serve in an extended tour of service with MARSOC. Once that tour is complete, they return to the usual assignments in their MOS.

SOCS-Ss receive a smaller amount of training than SOCSs. This training is geared towards 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. Applicants must also meet a variety of criteria including ASVAB, clearance, and medical screening criteria and passing the minimal physical fitness test (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 physical fitness test information is used during the selection process. However, MARSOC's training website advises that candidates should be able to demonstrate a 250 or higher PFT and maintain a 4 mph (15 minute mile) pace with a 45 pound rucksack regardless of distance prior to entering training. 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 CSOs 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 I is a three-week course that includes physical training involving 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 I for their ability to enter the 3-week Phase 2 course: Assessment and Selection. This phase is a highly competitive evaluation to identify which marines have what it takes to join MARSOC. 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.

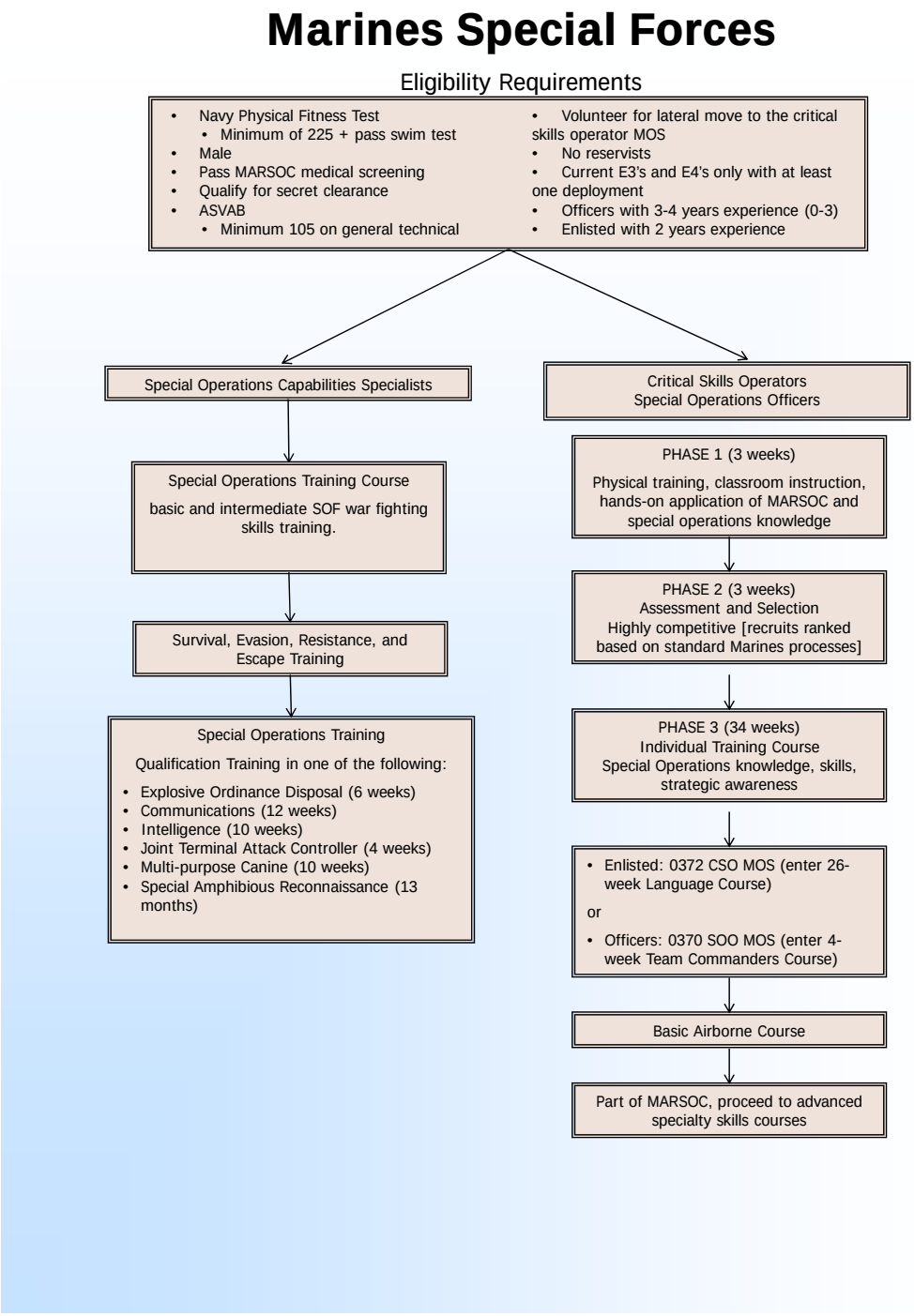
Marines who complete Phase 2 are selected to enter Phase 3, the 34-week Individual Training Course (ITC). Phase 3 is a nine-month course in which enlisted personnel (operators) and officers gain special operations knowledge, skills and strategic awareness. After the completion of the ITC course, enlisted 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 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 process for establishing standards

MARSOC has described their approach for addressing the WISR requirement as consisting of the following elements:

- Conducting a detailed job analysis for CSOs, SOOs, and SOCSs focused on identifying the tasks and abilities required on the job.
- Validating standards in the SOOs/CSOs assessment and selection course and the individual training and the SOCSs special operations training course.

Figure 7.1 Marines Screening Lifecycle



The following are the steps that MARSOC has outlined in their plan:

- Conduct a detailed job analyses for MARSOC positions of interest (SOOs, CSOs, and SOCSs).[9] 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 Individual Training Course 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 individual training courses, 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 describe the validation plans as including a hybrid of content-based validation and criterion-validation approaches, they have 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 are solely directed at validating the selection that occurs *during* the training courses. At the time in which our data collection ended, no plans were in place to validate the processes used to screen people prior to entering training.

Like USASOC, MARSOC has contracted OPM to execute their 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 review their work prior to completing our data collection efforts. However, OPM's statement of work was provided for us to review, and it serves as the basis for the description of the work provided below.

Job Analysis

OPM has identified several deliverables that will result from the job analysis methodology. The first two deliverables consist 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, etc., 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

[9] 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 they 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, they will use the regular infantry selection criteria once those are developed.

job incumbents performing job duties, interview incumbents and supervisors, and observe the way existing screening tests (if any) are implemented. The results of this process are then 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 is to hold SME panels (likely spanning two days) so that job incumbents can review and revise the list. Following that, additional SME panels with supervisors are convened to verify that the content of the list resulting from the incumbent SME panels is accurate. The list resulting from these SME processes is the second deliverable provided to MARSOC. It will also serve as the foundation for the next step, the online surveys.

Two online surveys will be produced and also delivered to MARSOC, one for supervisors and one for incumbents. Both groups will be asked to rate the tasks on importance and the competencies and physical abilities on importance, whether it is required for entry into the occupation, and the need for training in it. Incumbents will also be asked to rate the frequency of the tasks. Other scales may also be included. Analysis of the survey will be geared towards 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 is the fourth deliverable.

The list incorporating survey results will 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 will 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 will be delivered to MARSOC.

The very last deliverable resulting from the job analysis will be a report documenting all of the work described above.

Test Validation

As we discussed above, 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 during the early training stages, which experience a high dropout rate, and ability to perform the required job tasks required. The plans provided to us for OPM's validation step were far less specific than the plans for the job analysis. The validation plans acknowledge that both content and criterion-related validation strategies could be used and that the more evidence collected to support the content and criterion-related validity the better. However, they also note that the best strategy will 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 their statement of work.

If content validation is pursued it will include having a panel of testing and assessment experts review the tests and testing materials (including how they are administered and how scores are assigned and used) and the job analysis findings. The panel will also be asked to provide ratings on how well the tests represent the domain that the test is supposed to measure and how relevant they are for performance required in the job, based on the job analysis results.

OPM notes that if criterion-related validation is 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 meets specific statistical criteria would be needed. They plan to work with MARSOC to determine whether these requirements could be met. They also acknowledge that the existence of this 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.

Setting 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 they do 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 are 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 Step 1). MARSOC and OPM have taken steps to ensure that the personnel most knowledgeable about the job (i.e., job incumbents) are heavily involved, and they have 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 their statement of work are far less detailed. This makes it difficult to judge whether the results will provide sufficient support for MARSOC's selection processes. With respect to our recommended Step 3 (test validation), although OPM indicated they would take one of two acceptable approaches to validation (content validation and criterion-related validation), they did not provide any details about the approaches. We do not know what tests MARSOC or OPM would identify as relevant to validate (our recommended Step 2), nor do

we know what data might be obtained in support of the validity process or how that 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 recommended Step 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 minimums for the information resulting from the job analysis. There was no mention of how that information would be used to tie into the establishment of minimum standards (e.g., test score cut-off 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 lack of available documentation, there are large gaps in our understanding of the work that OPM is doing for MARSOC. When OPM provides documentation of the entire process and the findings, some or all of those gaps can likely be filled. However, as noted elsewhere, details 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 8. Navy Special Operations Forces

Five Navy occupations, known as the "Warrior Challenge" occupations, require physical screening tests to ensure personnel can meet the physical demands of the occupations. Only two of them—the Special Warfare Operator (SEAL) and Special Warfare Combatant-Craft Crewmen (SWCC, also known as Special Warfare Boat Operators) occupations—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[10] 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 Ordinance Disposal (EOD) Technician, Navy Diver (ND) and Aviation Rescue Swimmer (AIRR)—are already open to women and all currently have 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 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 physical screening test (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 Military Entrance Processing Station (MEPS).

At the MEPS, recruits complete a medical pre-screening report, take the ASVAB, provide documentation to demonstrate their eligibility to join the Navy (proof of age, 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

[10] 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.

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 12 1/2 minutes, the Navy reports that a 9-minute swim time is considered ideal. Similarly, 18 pull-ups and 90 pushups—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 whom they select to send to training. The physical fitness test scores are among the factors considered in making those selection decisions, and each occupation makes their 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 twelve weeks. 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 Qualifying Test (AFQT) and a set of combined ASVAB subtests. Candidates are also 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 1-4 (1 is lowest). Selection criteria for officers are similar to that of enlisted.[11] SEAL and SWCC eligibility requirements are summarized in Figures 8.1 and 8.2. respectively.

[11] SEALs 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 roles like department head or training officer in charge. Most senior roles include SEAL team commanding officers. There are typically 70-90 officer training slots. Successful training officers work in an SOF command that includes multiple services with at least 200 people. Criteria for SEALs Officer selection include proven leadership, strong language/cultural expertise, strong academic performance and highly-competitive PST scores.

Figure 8.1. Eligibility and Training Requirements for Navy SEALs

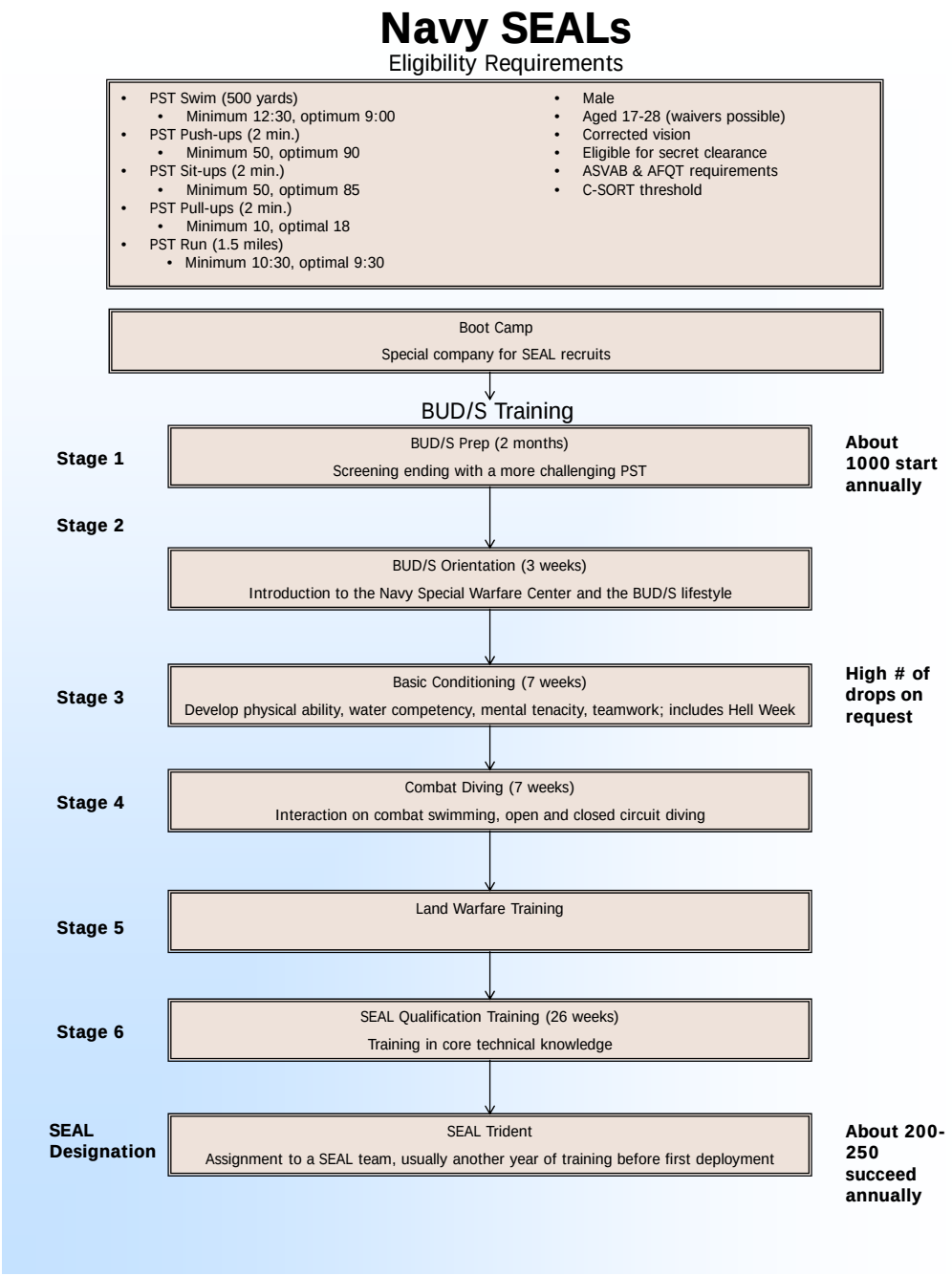
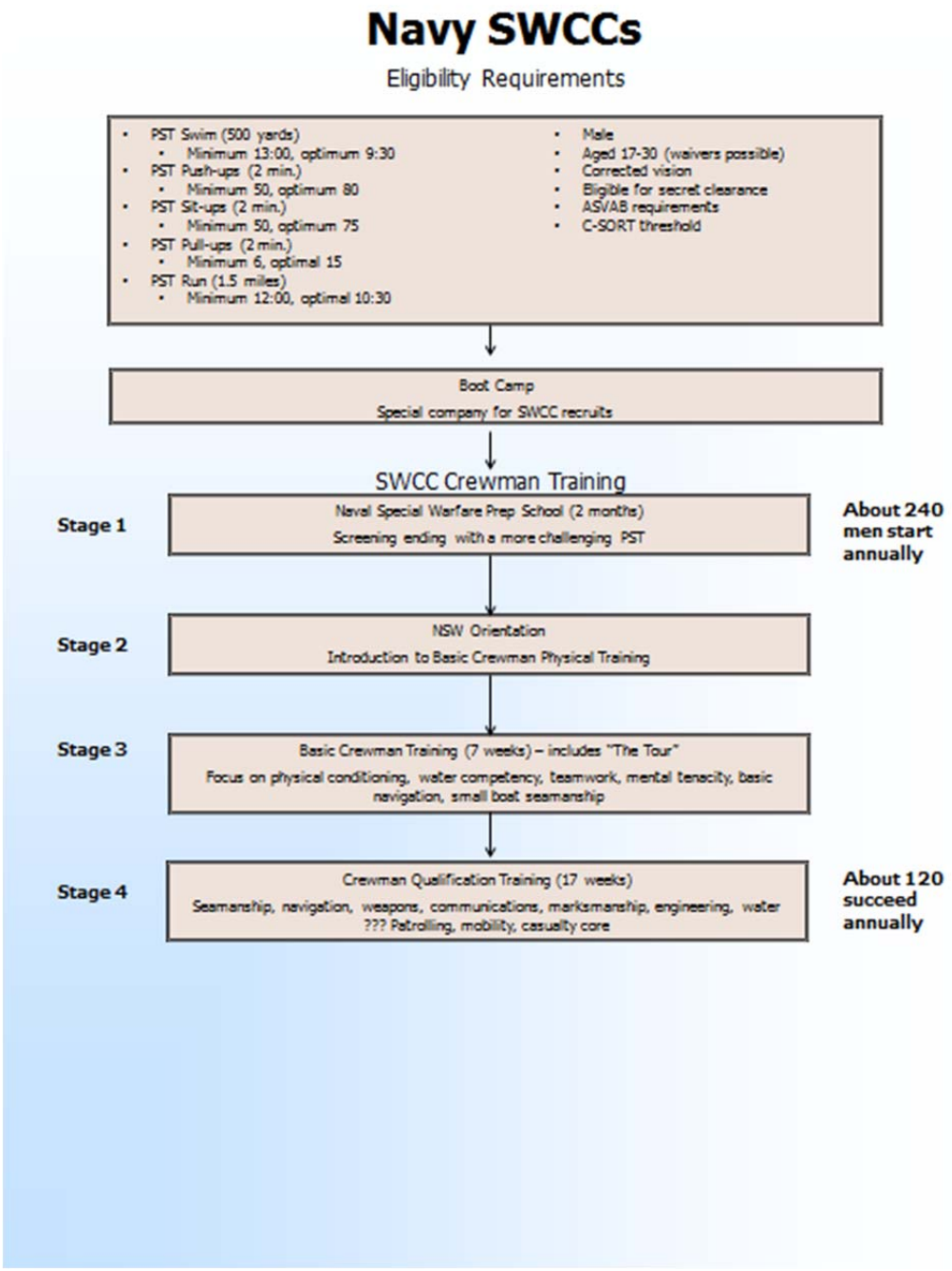


Figure 8.2. Eligibility and Training Requirements for Navy SWCCs



Because there are typically many more personnel interested in the jobs than there are spots in Basic Underwater Demolition/SEAL (BUD/S) training or Naval Special Warfare Prep School, only those judged to be the most competitive candidates on the basis of 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. Although we know that for both of these occupations PST results are combined using an established formula,[12] we do not know how the PST is combined with the other information to make a selection determination. We 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 the training itself is major selection hurdle for these occupations.

SEAL Training

Annually, approximately 1,000 personnel start the Basic Underwater Demolition/SEAL (BUD/S) training at the Naval Special Warfare Preparatory School in Great Lakes, Illinois, but only about 200–250 complete it.[13] 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 physical screening test 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* body weight dead lifts, an agility run (measured in seconds), a 300-yard shuttle run (measured in seconds), a three-mile run, and an 800 meter swim with fins. The ending physical screening test requires a 1,000-yard swim with fins in 20 minutes or under, at least 70

---

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

[13] 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.

pushups in two minutes, at least 10 pull-ups in two minutes, at least 60 curl-ups in two minutes, and a four-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 Naval Special Warfare 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 Naval Special Warfare 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 on each side) and a 1.5-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 tenacity and teamwork. Weekly tests include a four-mile timed run, a timed obstacle course, and a timed two-mile swim. "Hell Week," consisting of physical training for more than 20 hours a day with only four hours of sleep spread over five and a half 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 that occurs at BUD/S.

The second phase of BUD/S (Stage 4 of the overall training pipeline), "Combat Diving," is a seven-week course providing instruction on combat swimming and open and closed circuit diving, both of which are skills that are 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 freefall 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 8.5-month Basic Crewman training for SWCCs at the Naval Special Warfare (NSW) Preparatory School in Great Lakes, Illinois, but only about 120 complete it. Basic Crewman training has four stages.

The first stage, NSW Prep School, like the BUD/S Prep course, is a two-month screening training program that begins and ends with a more challenging physical screening test than the PST required for entry. The ending physical screening test during NSW Prep requires a 1,000-yard swim with fins in 22 minutes or under, at least 50 pushups in two minutes, at least 7 pull-ups in two minutes, at least 60 sit-ups in two minutes and a three-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, push-ups, 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 and mental intensity of the activities increases each week and culminates 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 Naval Special Warfare 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 SEAL and SWCC Selection Standards

The SWCC and SEAL occupations are under SOCOM oversight, but the Naval Special Warfare Command (WARCOM) has responsibility for validating standards and the work itself is being designed and carried out by researchers at the Naval Health Research Center (NHRC). The remaining sections in this chapter describe the evidence being gathered by NHRC to support the validity of the SEAL and SWCC process and the work being done to establish gender-neutral standards.

Evidence from Existing Studies

NHRC's data collection efforts do not include a re-examination of the PST requirements for screening to decide who can enter training in the closed occupations. NHRC instead directed us to existing research in over thirty 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 WARCOM personnel in 2014.[14] WARCOM explored PST and training attrition data from over 7,000 BUD/S students. The data show that only 27 percent of all trainees completed the

[14] 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 if it was conducted for other reasons.

training and that there is a relationship between PST composite scores and training success. These results suggest that the current 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—if, in fact, physical requirements are the primary reason that candidates drop out.

The researchers recommended that minimum PST standards for application to be a SEAL be increased 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 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 and 9:55 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 and both are therefore of direct concern in marshalling support for the use of the PST.

The study recommends implementation of a tiered minimum composite PST system, requiring a 1200 to enter into the SEAL challenge contract or delayed entry program, followed by a required 1100 to enter recruit training and enter NSW prep. Recruits would then need to score 1000 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). Currently, applicants who score 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 2012–2013 SEAL (n=291) and SWCC (n=86) training classes. The study explored the relationship between the three physical 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 both 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 Prep exit test and the NSW orientation assessment also show that completers performed better on every task than did the drops. This data supports 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.

Though none of the aforementioned studies included female participants, in 2013 the Navy Recruiting Command 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. Moreover, the data for females is somewhat inconsistent with information regarding the performance of females in the open occupations that was provided to us by representatives of those career fields.

Table 8.1. Test Scores of SEAL and SWCC Graduates and Non-Graduates

		SEAL		SWCC	
Test name	Test Elements	Non-Graduates	Graduates	Non-Graduates	Graduates
HPP Combine	Standing long jump (inches)	89.3	91.5	87.7	89.3
	Max pull-ups w/ 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 @ 1.5 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 2 attempts w/ 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
	800m swim w/ fins (hours)	0.59	0.58	0.60	0.58
NSW Prep	1000m swim w/ fins (hours)	0.73	0.72	0.75	0.72
	Max push-ups in 2 min	82.7	86.7	72.7	74.3
	Max sit-ups in 2 min	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
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
	1.5-mile run data for BUD/S students (hours)	0.41	0.40	0.43	0.41

Our evaluation of existing data for validating 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, that is, whether the role of PST scores and other factors in the rank ordering of applicants is the same for men and women and whether the rankings predict training success in the same way for men and women.

There are many gaps in the past research that still need to be filled to provide support for current use of PST scores for selection into training or physical screening test 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 males and females, something past studies do not explore.

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

Establishing Valid Gender-Neutral SEAL and SWCC Standards

As we indicated earlier, WARCOM turned to NHRC to conduct the research for establishing gender-neutral standards in response to lifting of DGCAR. In the documentation the NHRC provided to us and in our discussions they described the following data collection efforts which focus on investigating the extent to which SEAL and SWCC selection requirements that occur during training are related to occupational performance:

- Collecting data to update the list of operational and mission essential tasks to reflect current and anticipated future SEAL/SWCC mission demands
- Collecting data to evaluate graduates perceptions of the content of Hell Week
- Collecting data to show whether the standards expected in Hell Week accurately reflect operational demands of SEAL/SWCC missions
- Reliance on existing research to support the continued use of the PST

Job Analysis for SWCCs and SEALs

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

To develop the scenarios, NHRC held focus groups with 112 SEAL and 64 SWCC non-commissioned officers (participants were E-6s and above with an average of 6 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 that occur during typical missions that included details such 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[15] to finalize the scenarios.

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

---

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

- How physically difficult it is to perform the mission relative to all other missions
- How important it is 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 to all other missions

Our Evaluation

This job analysis uses methods that are consistent with the approach taken in the previous job analysis and are similar to approaches typically taken in collecting job analysis data. We do not know if respondents were asked if any important missions were missing, and we have not seen the results of the responses to these items or how they were analyzed.

The job analysis approach (Step 1) is consistent with the types of information collected in typical job analysis settings. Without seeing the data we cannot comment on many of the technical elements that would help lend support to the information including 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. The Navy was not able to share any of these specifics with us by the conclusion of our data collection period.

Identifying Physical and Personality Attributes Likely to Predict Performance on the Job

Other questions in the surveys described above 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 aforementioned mission sets. The attributes were a set that was 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 suggest in an early draft report that they plan to use this information to establish the content validity of the standards, but no explanation of how the data will be used to accomplish this was provided. Again, we have not seen the results of these survey items or how they will be analyzed.

Our Evaluation

This part of the Navy process aligns with Step 2 of our analytic framework. 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 could strengthen the conclusions. Even greater support for the conclusions could be established by collecting physiological measurements and conducting observations of people successfully performing the activity.

Linking Hell Week and The Tour to Performance on the Job

To establish this link NHRC included items in their survey asking SEAL job incumbents to provide their judgments as SMEs about Hell Week (the same process was used with SWCC job incumbents to validate 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, it appears that 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, the exact wording of the item and how the data from the survey was analyzed to arrive at these findings is unclear.

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

- Have you ever experienced a situation during an operation that was as challenging as Hell Week?
- How frequently during operations do you experience situations that are 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 are 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 conclude that Hell Week is valid, several of the past studies that we reviewed (dating back to the 1970s) raised a number of questions about whether Hell Week content is really justified. Some of the studies noted that the difficulty of Hell Week varies

from class to class and that instructors are idiosyncratic in how they treat each student, with some students appearing to be challenged more than others. Many of the studies noted the extremely low pass rates, even among the most physically fit and prepared candidates. Some have asked whether many of those who were driven to quit would actually have been successful on the job. Multiple reports recommend standardizing the training difficulty and making other adjustments to reduce attrition. Senior leaders in the Navy and researchers alike are on record as having questioned the need for Hell Week over the years, well before DGCAR was ever lifted. We are unable to tell what changes have been made, if any, in response to these past studies. It does not appear, however, that NHRC has explored how consistent the training difficulty is from person to person and class to class. Therefore, these are still areas worthy of further investigation.

Our Evaluation

The data supporting the link between the content in training and the content on the job is entirely based on perceptions of the links by job incumbents in responses to a few narrow items on a survey. Job incumbents believe Hell Week helped them gain confidence in themselves and others, they believe it is essential to preparing a candidate for operational performance, and they have faced equally difficult situations on the job. Although this is evidence that can be marshalled 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 report having faced an equally challenging situation to Hell Week on a mission, it is not clear that similarity in judged difficulty alone is sufficient to show that the Hell Week content is reflective of the content on the job. For example, it is plausible that someone could judge learning to figure skate to be equally as challenging as learning to play the piano; however, teaching you to figure skate or screening you on your figure skating success would not help you to be a successful pianist.

Similarly, 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, ways that might not be accompanied by attrition of large numbers of personnel. Although few would argue that confidence would be irrelevant to successful performance in these occupations, it is not clear that such confidence would be broken if the content of Hell Week were different. That would be a necessary condition to justify the high-levels of attrition that occurs during Hell Week. It would be needed to further support the content validation of Hell Week.

Lastly, given the past concerns about the variability in the difficulty of the Hell Week training and the high attrition even among those who had competitive entry scores on the PST and other screening criteria, reasonable questions regarding the validity of Hell Week could still be raised.

Linking Hell Week to the PST and Physical Testing Earlier in Training

In the preliminary findings provided to us, NHRC presented correlations exploring whether various aspects of SEAL and SWCC training and testing are related. However, what type of data was used to create the correlations reported was unclear. 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. On the other hand, it is possible that scores obtained by actual trainees on each of the elements are what was used to compute the correlations. The researchers conclude on the basis of the data that SEAL and SWCC training outcomes do reflect physical testing scores. The intent of this analysis appears to be to show that performance in earlier training events is related to performance in later selection events as a justification for using the earlier events to screen people out of training long before they begin Hell Week or the Tour.

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

Without knowing more about how the correlations were computed and other specifics about the data we cannot provide a complete assessment of its usefulness for establishing the validity of the physical testing (Step 3 of our framework). This is an example of why detailed and complete documentation of the work is vital. Nevertheless, it appears that there are potential gaps in the logic of 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 point in time demonstrates that they are reliable at rank-ordering trainees on the activities tested, but it does not prove that that this rank ordering predicts how well the trainees would do on the job. There may well be a relationship, but that relationship cannot alone be used to support the validity of the test.

In addition, if some of the data are judgment based (where job incumbents have judged there to be relationships among the testing and Hell Week performance) rather than based on actual scores, it is possible that those judging the relationships are simply wrong in some cases. No evidence to support the accuracy of judgments was provided.

NHRC's Overall Conclusions Regarding Validation of the Standards

Based on the efforts summarized above, NHRC researchers concluded in their preliminary findings that:

- Physical standards are reflective of the physical occupational demands for both SEALS and SWCC.
- First phase of SEAL and SWCC training reflects physical testing administered earlier in training.
- The evaluation tasks administered during training reflect the desired characteristics of SEAL and SWCC operators (such as maturity, integrity, humility, etc.,).
- Hell Week activities are essential and useful for preparing recruits for operational performance.
- Situations as difficult as Hell Week occur during operations.
- Completing Hell Week increased the confidence SEALS/SWCCs have in themselves and in their teammates.

Our Overall Evaluation

Selection into the SEAL and SWCC occupations is begins when applicants are initially selected for training and continues through the training period. Applicants must have minimum PST scores, but initial 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. Almost all of those who continue beyond Hell Week will graduate and enter the occupations, but a high fraction of those initially selected drop out of training up to and during Hell Week.

Because Hell Week serves as the final selection point and the point at which many trainees drop out, the NHRC research for WARCOM focuses whether the activities during Hell Week reflect mission requirements on the job and whether scores on the physical assessments during training predict performance during Hell Week and other training activities. When the researchers conclude that SEAL/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 predicts performance on the job. Such data

would go a long way towards furthering support for the continued use of Hell Week as it now stands.

- We cannot evaluate the analysis done to determine whether the sequence of physical assessment tests predict training performance and completion without more documentation of the methods and data used.
- None of the data collected by NHRC included females because there are no women currently on the job or in training. It is therefore unclear whether the relationships reported between earlier training testing performance and performance at 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 (chosen based on scores on the first more challenging physical tests in the prep courses, for example).
- 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 performance during Hell Week, during the other training blocks, or on the initial PST prior to entry into training (i.e., Step 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 is based on an early draft of the write-up of the methodology for the NHRC work. That write-up lacks the detail and clarity necessary to fully understand and critically evaluate the approach they have taken to validating the work. Much of it is focused on discussing the importance of presenting a persuasive argument for validity instead of on the actual work conducted to support that argument. It does elaborate on some of the data collection efforts, but again, the details and rationale for the processes are incomplete. Fully documenting the work will be key to determining whether it will stand up to scrutiny. Once such documentation is in place, the work should be reexamined. Some of the potential gaps we have identified here may disappear with a more detailed and complete description of the research and how it is ultimately used to set valid, gender-neutral standards.

Chapter 9. Air Force Battlefield Airmen

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

- 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 themselves are quite different. Combat controllers oversee air traffic control in austere environments. Special operations weather team personnel serve as meteorologists who provide intelligence to inform mission planning, route forecasts, and special reconnaissance. Tactical air control party airmen coordinate air support of ground combat and clearing of airspace. Pararescue are trained as emergency medical technicians to operate in humanitarian and combat contexts on conventional and unconventional rescues by air, land, and sea. Combat rescue officers also coordinate and directly engage in rescuing people and resources. Special tactics officers 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 AFSs together accounted for 4,686 positions closed to women (Air Force High Level Implementation Plan on Gender Integration, 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 a occupation happen during basic training using data from the MEPS to determine eligibility. AFI 36-2101 (2013) governs the process for classifying Air Force officers and enlisted into their respective occupations.

The majority of officer AFSs 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 strength aptitude test (SAT). The strength aptitude test (SAT) has been in place for decades, and the PAST test 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 strength aptitude test 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. In addition to being male, these conditions include (see Figure 10.1): 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); normal color vision; at least 20/70 vision in both eyes, correctable to 20/20; passing the standard military physical and the Class III Flight Physical; a minimum height of 4'10" and maximum height of 6'8"; and no more than 250 pounds (the maximum weight for jump school). Officers face similar eligibility conditions for the battlefield airman occupations.

The Strength Aptitude Test (SAT)

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. Many enlisted occupations, however, 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 less than 100 are eligible for fewer occupations. For a complete list of the lift requirements for every AFS in the Air Force, see the Air Force Enlisted Classification Directory (AFECD, 2013).

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 over 100).

Given this, many women and most men achieve scores that meet or exceed the battlefield airman SAT minimums. For more on the strength aptitude test, see Sims, Hardison, Lytell, Robyn, and Wong (2014).

The Physical Ability and Stamina Test (PAST) Test

The PAST has been used for several years to pre-screen 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 that are already open to women.[16] The PAST includes a 25-meter underwater swim (assessed as either pass or fail), a timed 500-meter surface swim, a timed 1.5-mile run, a count of chin-ups completed in one minute, 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 - 500 meter swim	Max time - 1.5 mile run	Min Pull ups	Min Sit ups	Min Push ups
Pararescue Jumper (PJ)	Closed	Pass/Fail	10:07	9:47	10	54	52
Combat Control Team (CCT)	Closed	Pass/Fail	11:42	10:10	8	48	48
Special Ops Weather (SOWT)	Closed	Pass/Fail	14:00	10:10	8	48	48
Tactical Air Control (TACP)	Closed	NA	NA	10:47	6	48	40
Explosive Ordinance Disposal (EOD)*	Open	NA	NA	11:00	3	50	35
Survive Escape Resist Evade Trainer (SERE)*	Open	NA	10:00 (200 meter swim)	11:00	8	48	48

*EOD and SERE are already open to women.

For enlisted applicants, the PAST test 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

[16] For more specifics on the PAST, see AFOCD and AFECD, 2013.

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.

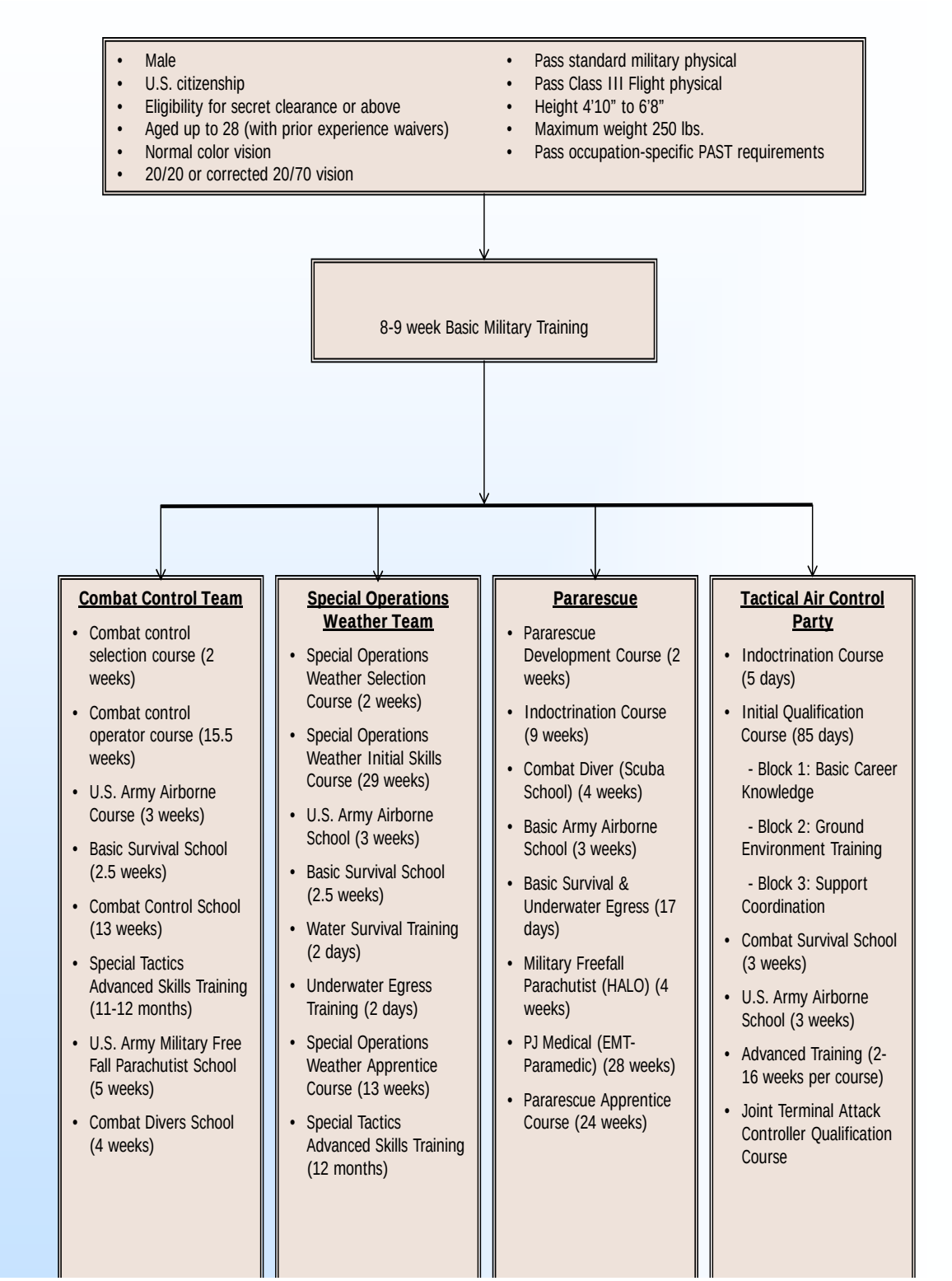
For STOs and CROs (the battlefield airman officer jobs) a modified version of the PAST (with slight variations in the content and ordering of the test) is administered as part of the initial application for entry into training. For example, the STO test requires one 25-meter swim instead of two and it requires a 1500-meter swim rather than a 500-meter swim. For those events that are identical to the enlisted PAST (such as pull-ups), the minimum scores differ for officers.

Training and Continuation Requirements

All enlisted recruits attend basic training and then proceed into occupation-specific technical training. Each technical training pipeline is unique, but most are divided into multiple blocks of training each intended to address different elements of the job. During and upon completion of the blocks, trainees are required to demonstrate proficiency in order to move on to a subsequent block. Those who do not meet minimum expectations for completing a block 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 washout and wash back trainees. As a result, training itself 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). The battlefield airmen occupations all include some sort of assessment and screening block that is 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 that are currently 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 outlined in AFI 13-219 (volume 2) 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 (see the AFECD and AFOCD). Lastly, a new operator test for battlefield airman occupations has recently been established to ensure the standards required during training are also being maintained in the operational force.

Figure 9.1. Eligibility and Training Requirements for Enlisted Battlefield Airmen



Establishing Occupational Entry 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 the now-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 on the force-wide fitness assessment program and policy. The unit is led by a civilian exercise scientist and includes seven other personnel dedicated to the effort to establish battlefield airman standards: two part-time officer aerospace physiologists, three full-time research assistants, and two AETC/A5/8 Studies and Analysis Squadron analysts. Although the unit is conducting many elements of the research and data collection effort, the Air Force also commissioned an FY2014 RAND Project AIR FORCE study for some elements of the research. The work to develop new battlefield airman standards began in 2011.

The Air Force is currently exploring new physical aptitude tests to screen people for entry into battlefield airman specialty training pipelines and ultimately replace the PAST. The result will be a new set of occupation-specific physical screening criteria (referred to as Tier 2 fitness standards).[17] The following sections describe the process the Air Force has underway to identify and validate the new battlefield airman screening criteria. The information described below was gleaned largely from our interviews with the RAND Project AIR FORCE 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 non-commissioned 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.[18] SMEs were sampled

---

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

[18] 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. They develop the task lists by first soliciting SME input to confirm the relevance of the existing tasks from prior year's lists and identify gaps. Then they survey all personnel and ask them to rate each task that was confirmed or added by the SMEs. Task lists for officer occupations are only developed on special request, but the occupational analysis flight follows similar procedures for developing them when they are requested. Task lists resulting from the occupational analysis flight's

to ensure representation within each job by mission, unit, and environment and were required to have had at least one operational deployment within the recent 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 describe the physical actions used during the activity (pull, press/push, bend, squat, lift, crawl, climb etc.) They were also asked to provide relevant details like 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 specialty 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.

Criterion-Related Validation Study to Replace the PAST

Data collection for the Air Force's criterion-related validation effort was slated to begin April of 2015. In that data collection effort, the Air Force identified and administered a range of physical tests for use as potential predictors and they designed and administered a range of physically demanding job performance simulations for use as measures of performance on the job. They estimated that, when complete, the tests and simulations would be administered to a sample of at least 200 personnel—50 job incumbents (all male) and 150 tech training students from other careers (including about 80 females). 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 would be introduced to the simulations and tests and provided training and practice opportunities to ensure they are familiar with the activities and know what is expected of them before being tested. 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, lack of variance in participant scores), refine key test administration features (such as testing times, distances of rucks, height of walls, repetitions, and appropriate weight loads), verify that key activities were still judged as appropriate and realistic

---

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 above.

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 Physical Screening Tests

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 non-military 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 greatest 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 pushups). Using the results of the pilot test, the number of tests was further narrowed. Only a subset of the original 60 were 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 March of 2015, but they were expected to include a variety of different types of screening tools, with slight variants on each type.

The Simulations

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 are time to complete each activity and/or total distance completed within an allotted timeframe.

Some of the simulations are isolated and short in duration. For example, one is intended to simulate a boat carry over obstacles across a beach area before putting the boat back in the water. Participants pick up a bag with a handle (meant to simulate the boat) and walk through pea gravel the distance of a typical beachhead. Another simulates a rock climb by having participants climb up a wall and then pull up their rucksack. Others, however, will be 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, is the most complex and time consuming of the simulations. It starts with a five-kilometer ruck march after which participants complete the following activities in sequence: a low crawl 24 inches high (Army standards), a buddy drag, an evasion maneuver with cones and obstacles (various heights walls), a maneuver over an eight-foot wall with a two-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 (Florida, Alaska, Hawaii, Colorado, etc.,) 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. From the data collected at all of the full mission profile training locations researchers identified key design features for the simulations.

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

How the Criterion Valuation Results Will Be Used

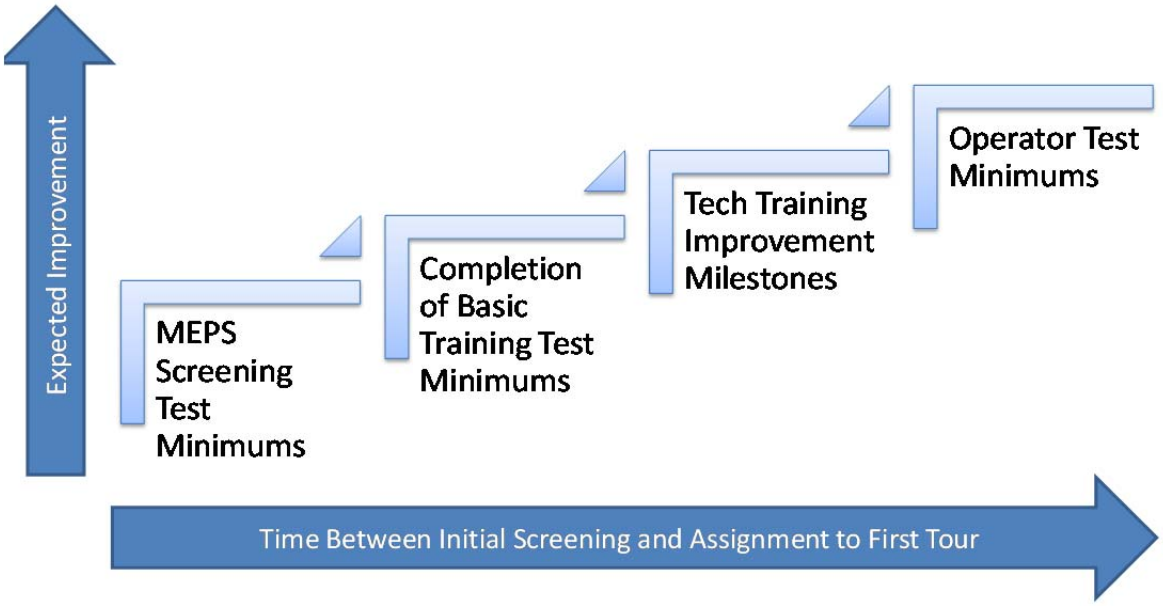
The criterion-related validation study data will first be used to establish the recommended annual testing standards (such as how many pull-ups) for the battlefield airman operators (the people currently performing the job). In other words, the study aims to identify the tests and minimum scores that best determine who is physically ready to perform on the job and who is not. The resulting minimum scores on those tests (specific to each occupation) would be used to certify each current battlefield airman's capability annually. Once the tests and score minimums have been determined, they will be proposed to leadership for concurrence.

After the operator tests and minimum scores have been established, the Air Force plans to work backward over the training pipeline to establish the training entry requirements. As shown in Figure 9.2, entry requirements used to qualify personnel for training will be designed to account for improvement and development over the training period. How much improvement to expect will be estimated using data from published research on physical aptitude training and archival data on male trainees in past battlefield airman training pipelines. Working backwards from the desired performance end-state (i.e., minimum scores on the operator test), the analysts will use the 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 will 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 pararescue, 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 will necessarily be lower. As a result,

the researchers have acknowledged that having different training entry standards for two different career fields even if the final physical demands on both occupations are the same is a very real possibility.

Figure 9.2. Incrementally Higher Minimums Account for Improvement Gained From Training



How the Operator Test Minimums Will Be Established

The researchers first set minimum standards for performance on the simulation activities. To establish those minimums, a subset of the 50 criterion validation study participants who are experienced operators participate as SMEs in a standard setting panel. The panel takes place during the criterion validation study itself. Those SMEs selected for the panel (i.e., those identified by the career field as having the appropriate experience and expertise) first complete a simulation themselves as part of the normal study data collection. After completing the simulation, they are told how they performed on it (e.g., how fast they completed it) as part of their role on the panel. They are then be 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 are shown to leaders and compared to actual operator performance on the simulations as a final check on the accuracy of the minimum performance levels established by the panel participants. Once minimums on the simulations are identified, minimum scores on the tests are determined using the statistical correlation between the test scores and the simulation activities on a subset of the participant sample. The remainder of the sample is used as a hold-out sample to cross-validate the minimums. That is, the statistically derived test minimums are applied to the hold-out sample and the amount of error in

predicting who will be successful in the simulations is explored. Test minimums may be revised depending on the results.

Once the operator tests are selected and minimums are set using the process outlined above, the researchers plan 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 is intended to allow Air Combat Command and AFSOC to verify that the established standards are working 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 perform 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 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 will be used to define the operator test minimums. As explained above, the test scores required at earlier points in the career (e.g., during training, or upon entry to the service) will 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 the Step 1 of our 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 have not seen a write-up of the findings from the job analysis data collection process, so we cannot determine whether the data analyses and conclusions drawn by the researchers are 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 Step 2. They used a well-reasoned and systematic process for initially narrowing the list of potential tests 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 focuses on using empirical findings to drive the final test content, which again is consistent with recommended practice.

With respect to our recommended Step 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 their 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 cannot confirm how well the simulation activities actually 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 details that are important in evaluating the final results.

Other important factors to consider with respect to our recommended Step 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 to explore whether the test scores demonstrate the appropriate statistical properties (including sufficient variance) will be needed. The resulting regression findings and the conclusions the researchers draw with respect to those findings are also important.

With respect to our recommended Step 4, the Air Force has articulated a plan for how they will establish minimum scores on the tests that is also consistent with recommended practice. They first plan 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 steps proposed to check the SME judgment against the judgment of others knowledgeable about the career field is a strength of the methodology; however, if the double checking by others suggests changes are needed, there will need to be a strong rationale and sound justification for those changes.

The next step is 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. There are many decisions to be made statistically crosswalk the data that could lead the results to be faulty.

Lastly, the Air Force also plans to use estimated training gains over time to establish the minimum test scores required to initially qualify for training. The data on which they plan to base their estimates for training gains has not been provided to us, so we cannot evaluate it.

In sum, it appears that many elements of the Air Force's criterion-validation effort are consistent with recommended practice. As explained above, the researchers have taken steps to collect solid data on which to base their decisions at important points in the validation process, and they plan to have data supporting many of the important links that are critical in a well-designed criterion-validation study. However, the formal write-up of the methods, analyses and the findings are still forthcoming and therefore many of the details of their data analysis decisions are still unknown to us. In addition, although there are many strengths to the approach that can lend credibility and support to any resulting test score minimums, there are some

potential gaps in the work. In particular, examination of bias of the testing by gender is one area that was not addressed in the plans described to us.

Chapter 10. Conclusions

Comparing Across the Services Efforts

Each service took a slightly different approach to amassing evidence to develop and support their screening standards. Differences in their approaches should not be taken to 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 below.

Operationalizing "Physical Screening"

Each service conceives of their 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 most heavily on validating the training content. However, in each case, the information obtained through the research is useful for informing the validity of the other screening elements. For example, the Army designed a simulation-based criterion validation study in which individual-level task simulations were designed, measured, and analyzed with attention to detail and data linking them to more realistic occupational task requirements. The Marine Corps undertook two studies, one designed to modify its fitness test to 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. The Air Force's work validating the initial screening criteria included an especially thorough job analysis which helps clearly define the requisite training content. The Navy focused on the training content because that is where much of the intensive screening takes place through both voluntary and involuntary attrition. The Army and Marines Special Forces work by OPM can likely inform whether the content of both the training and the screening tools is relevant.

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 across the Army and Marines efforts. Because the two services have taken very different approaches to establishing the standards for infantry, it will not be surprising if

they end up with somewhat different screening criteria. The differences may not necessarily mean that either would invalidate the screening process; that is, both could be equally valid. However, if the resulting differences 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 those differences will likely need to be reconciled with attention to the legitimate reasons for those differences to exist. For example, if the screening process in one service happens much earlier in someone's career than in the other, a lower set of physical screening minimums could be justified as there would be more time for personnel to train to 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 are slightly different from Army requirements, thus justifying the 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 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 are utilized. TECOM motivated its approach by the observation that all members of the combat forces (regardless of specialty) must be capable of meeting the physical demands of any combat arms occupation. This anticipates that these Marines will be called upon to perform duties 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 occupation only. This too is an area where differences will be apparent and warrant justification.

Table 10.1. Summary of Key Features of the Service Approaches

Service	Selection Process Being Validated	Step 1 Job Analysis	Step 2 Identifying Screening Criteria	Step 3 Validation
Army Combat Arms	Screening before training	Review of existing job-analysis materials through 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 OPM using occupational information, site visits, job incumbent survey	Current training activities	Content validity, details to be determined
Marine Corps Combat Arms (phase-1 study)	Screening before training	Job tasks identified from current training and readiness manuals, which rely on occupation-specific task lists regularly updated based on SME review and a job incumbent survey	Elements of current Physical Fitness Test and Combat Fitness Test	Concurrent criterion-related validation to determine how well candidate tests predict performance on basic physical tests roughly similar to physically demanding job tasks
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 Forces	Training	New in-depth job analysis by OPM using occupational information, site visits, job incumbent survey	Current 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	Current 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
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

Unavoidable Limitations in What Can Be Completed Prior to Opening Positions

As noted in our description of best practice methods, no single research effort can address all issues, and no research study is without weaknesses and gaps. The gaps we were able to identify given the stage 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, Step 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. Three gaps in particular are issues common to all of the services' work in support of standards for the closed occupations.

No Existing Female Applicants, Trainees, and Job Incumbents

No women are in the closed jobs yet. As a result, there was not a pool of incumbent women for the researchers to draw upon as participants in the research. The only subject matter experts with deployed experience performing the job are male. Women participating in simulation activities (such as in the Marine Corps, Army and Air Force efforts) do not have operational experience comparable to the male counterparts. This omission of job-experienced females poses an unavoidable dilemma. Because the NDAA has mandated that evidence supporting the validity of the standards be in place prior to opening the jobs to women, this means that no women will have experience in the job prior to the positions being opened. As a result, a major limitation of any standards the services establish is the inability 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 will likely change as people adjust to the opening of the positions and it becomes a more accepted career path for women, and as women become interested they will undoubtedly begin to prepare in earnest to meet the physical demands. As a result, we strongly 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[19] to include examination of validity within each

[19] How frequently each stage should be repeated depends. 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 conducting 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

gender group will be essential to continuing to justify continuing use of the tests and selection criteria. This is discussed further in the section crosscutting issues below.

Unforeseen Impacts of Implementation of Testing

Implementation of the testing (Stage 5) itself can lead to unforeseen changes in the validity of the testing. This could apply regardless of gender (i.e., validity for both men and women applicants could be affected), and it is something the services will need to watch closely. Collecting data on this in the months and years after establishing the standards will be important for ensuring that the tests and criteria perform as expected.

Research Today May Fully Support Implementing the Standards, But Future Research May Still Show Changes Are Needed

The services research efforts are intended to establish standards on the basis of the evidence amassed so far, but more research ultimately will be needed to fully determine whether how well the tests and test minimums are working (this is explained in Stage 6). It is possible that standards may need to be adjusted up if there are injuries or failures to perform adequately among those who meet the standards. Alternatively, if evaluation of the testing shows that too many people who would be successful are being screened out, the minimums might need to be adjusted downward to admit more qualified personnel. Lastly, future research may find that other tests do a better job of screening personnel with less adverse impact against gender and race groups. In those cases, changing the tests entirely may be needed.

We fully expect that as additional information is amassed and the available tests evolve, the services will need to make adjustments and refinements to the selection processes. Such adjustments should not alone be taken to mean that the work prior to opening the occupations was inadequate or faulty. Instead, if those adjustments are informed by new sound data obtained after implementation, it will be one good indicator that the services are continuing their investigation and have adopted the underlying spirit of the work: to continually seek more accurate ways to select their personnel.

---

verify that the key factors (such as length of the intervening training, administration procedures, test difficulty, and individual preparation) have not led to unforeseen changes in validity and 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.

Other Crosscutting Issues

Formal Documentation of All Aspects of the Work Is Needed

All too often the military conducts research to support their policy decisions, but fails to retain detailed documentation of the work after a policy decision has been made. Unfortunately, when not documented properly, the work holds little value in lending support to the practices over time. Details such as the overall statistical and methodological approaches, summary statistics, data analyses, sampling approach and participant characteristics, etc. are all necessary for experts to be able to judge the soundness of the research findings. Without those details, evidence to refute any challenge to the selection practices ceases to exist. For that reason, we strongly recommend that the services create and retain detailed write-ups of all research conducted to support and evaluate occupational physical standards.

We also advise making the documentation available to the public or, at a minimum, permanently available to personnel in the services in need of the information and experts in personnel selection representing those personnel. Making the work public ensures it can be found easily through online searches. Leaving it unpublished as an internal report runs the risk that it could be lost or forgotten over time. For that reason, we strongly advise that if it is not published, it instead be submitted as an official internal publication accessible and searchable by anyone in the service (e.g., at the Defense Technical Information Center). Publication of the work or making it available to all service members would demonstrate transparency and garner buy-in from service members and the public.

At the time at which we completed our data collection, each of the services had various stages of documentation underway. The Secretary of Defense requested documentation by prior to October 1, 2015, before decisions on opening occupations to women and implementation of the standards. The existence of such documentation prior to implementing the standards can help ensure that inquiries from the public are met with clear, consistent, and accurate information regarding the work.

Process for Establishing Minimum Acceptable Scores Still Needs to Be Reviewed

At the point of completing our research, the services had not yet established minimum selection standards. However, this step 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 impact 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 chose 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 too few people qualify. On the other hand, if the standards are set too low, the mission could also suffer as 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 step in the process will be critical to supporting the use of the screening minimums chosen.

The Implementation Step Still Needs to Be Investigated

The implementation step 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 how 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.

Research Needs to Continue After the Standards Are Implemented

Although the services efforts have all been huge undertakings, not all research can be done a priori. More research will be needed over time. It will be important to follow up after implementing the standards to see if the standards have good predictive validity in practice. Essentially, how good are they at distinguishing the good performers from the inadequate ones. However, the ability to explore such relationships later due 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 selected 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 that are already in place. These issues would need to be accounted for when examining how tests are functioning after they have been implemented. 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 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). OSD should therefore 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.

Final Thoughts

The call to develop valid standards has been taken very seriously by the services. All of the services have dedicated a large amount of time and resources to their work in response to the lifting of DGCAR. As a result, the service efforts have been very large undertakings. Some have involved setting aside dedicated testing locations, simulation equipment, and scientific physiological measurement equipment. All have sought to involve personnel with the appropriate research background and expertise. Some services had the requisite experts in house, whereas others sought out the assistance of experts outside of their organization.

The numbers of voluntary participants joining in in the work have also been impressive. Calls for participants (both male and female) 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 have put forth reflects a valiant effort to accomplish exactly what was being requested: the establishment of gender-neutral valid physical standards.

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 steps to establishing physical standards (Hardison, Hosek, and Bird, 2013) 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 towards 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:

- 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 Department of Labor have adopted these sources and so it would make sense for DoD to adopt them. As such, the definitions we provide in this Appendix are intended to be consistent with those endorsed in the documents cited above.

These authoritative sources are intended for a practitioner audience that already has some technical understanding of personnel selection but they are less accessible to other audiences. They also do not address all the specific terminology issues currently facing the DoD community. This Appendix is therefore intended to supplement the information provided by those sources by offering a non-technical 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 the sources cited above.

This appendix is intended to promote a common language among the researchers involved in setting occupational physical standards in the military, it is also intended to help ensure that DoD policymakers, Congress, and the public understand what the services mean when they describe the work they have done 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. 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. These 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 current 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. 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 knowledge, skills, abilities (KSAs). Any criterion that is used to exclude or disqualify someone from a job is essentially operating as a test or a measure of their 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 examples provided above, 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 that 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.[20]

---

[20] 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 subject matter experts 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.

Used interchangeably with the terms *hurdle, cut score* and *requirement*, in personnel selection the term *standard* 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 the importance of maintaining standards, this is an important point to stress. 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 that are 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.

Force-wide 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 healthcare 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 force-wide.

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 Fiscal Year 2014 National Defense Authorization Act, 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.

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, in practice the term gender neutral often can be confusing. 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 that are 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. And both should therefore 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 EEOC and the 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.

Here, 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*. *Predictive bias* 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 under predict 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 against 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.[21] 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 whether someone can lift 80 pounds repeatedly to chest height will likely predict 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 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 only occur 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 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 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

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

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 subject matters experts 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 etc. believe the test looks like or 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 that is 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 not documented, then as researchers leave, retire, or forget what was done, the institutional knowledge of the work 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 SIOP and the APA 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 by 2016. The Service efforts we are tracking focus on occupational-classification standards and as the first women move through the pipeline in the newly opened occupations, additional effort will be needed to ensure that validated standards are applied at later stages.
- Gender-neutral standards are not just standards that are the same for men and women. To be valid, they must also be unbiased (or fair)—exhibit the same relationship to job performance for men and 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 Already Open to Women

This chapter focuses on the services' efforts to establish valid physical standards for occupations that are already open to women. 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 a clearly designated occupations they have identified as physically demanding and they both have established a standardized physical screening process for those occupations. We therefore discuss the Air Force's and the Navy's screening criteria and their efforts to validate those criteria in detail later in this chapter. But 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 that are 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 are in existence 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 (31B) names the following physical job requirements (DA PAM 611-21, Table 10-31B-1):

1. 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)
2. Frequently lifts 42 pounds over head
3. Occasionally walks slowly for 2 out of 6 hours while carrying 170.9 pounds
4. Frequently stands for extended periods of time.

Each MOS is also assigned a physical demands rating according to the following scale (Hollander et al., 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 DA PAM 611-21 (section 10-31B) as "heavy."

This is just one illustration of an open occupation that is known to be physically demanding but for which there is no systematic physical screening test 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 below), 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 by having no standards it 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 are high as well. 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 underway.

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 4, is administered at the MEPS and is used to qualify enlisted applicants for certain AFSs. 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 AFSs 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 lbs

(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 Air Force Enlisted Classification Directory (AFECD, 2013).

Many of the minimum scores for the SAT were established when it was originally instituted in the 1980s. Job-specific minimum scores are only adjusted in response to a direct request for re-evaluation by the career field managers. 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). That data is 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 that data is scarce.[22]

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 has 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, they will consider the use of new any new tests recommended to improve screening, as a result of that study.

Although the Air Force has a comprehensive effort underway 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.

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—Explosive Ordinance Disposal Technicians (EODs), Navy Divers, and Aviation Rescue Swimmers (AIRR)—are already open to women and are discussed here. All three are considered physically demanding by the Navy, have physical screening requirements for training eligibility, have physically demanding training

[22] For more background on the SAT, how it was developed and how minimums have been established and revised over the years, see Sims, Hardison, Lytell, Robyn, and Wong (2014).

elements, and have 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 10.1-10.3 below. Minimum PST scores for each occupation are shown in Table 10.1. Although there are clearly defined minimum scores, each career field also has identified ideal scores that are needed for an applicant to be considered competitive. The ideal scores are much more stringent than the minimums. For example, for both EOD and Navy divers, the 500-yard swim and the 1.5-mile run have a maximum time of 12:30 minutes, but optimum times of 9:00 and 9:30 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	Diver	AIRR
Swim 500 yards - breaststroke or sidestroke [in minutes]	12:30	12:30	12:00*
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 1.5 miles [in minutes]	12:30	12:30	12:00

SOURCE: Navy.com (2015) http://www.navy.com/careers/special-operations/air-rescue.html#ft-training-&-advancement
*AIRRmay use sidestroke or breaststroke and utilize American crawl/freestyle or a combination of all.

Figure B.1 Navy EOD Screening Lifecycle

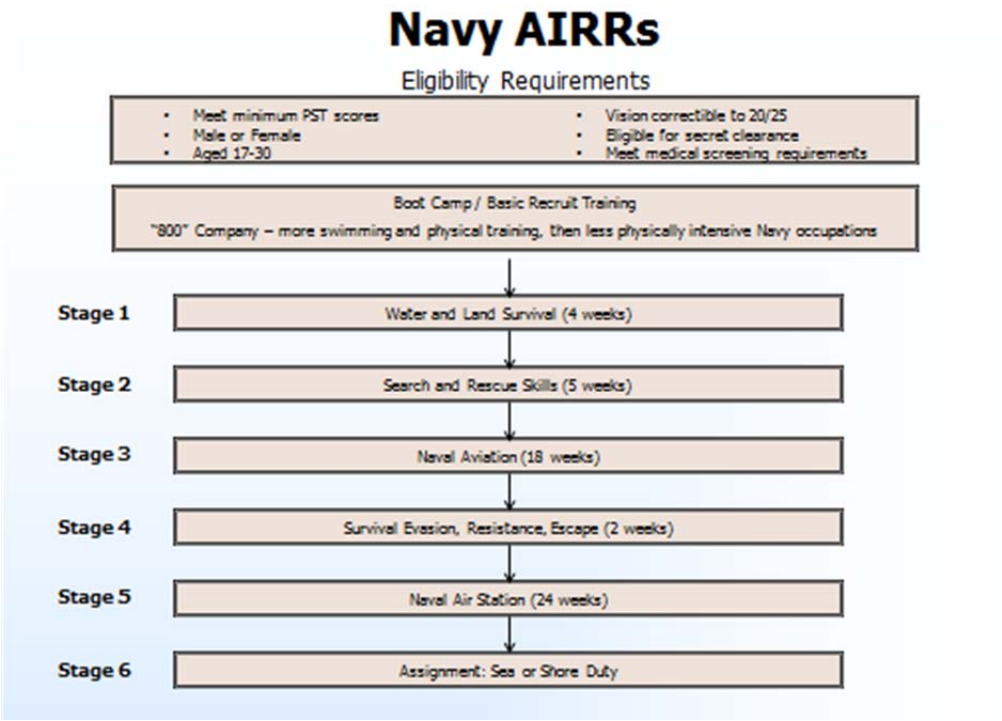


Figure B.2. Navy Diver Screening Lifecycle

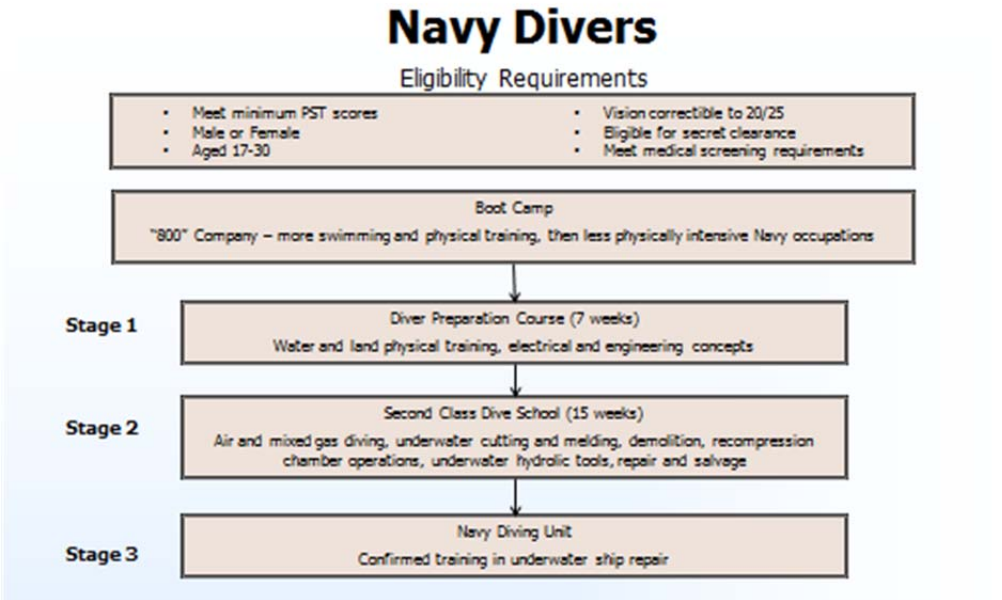
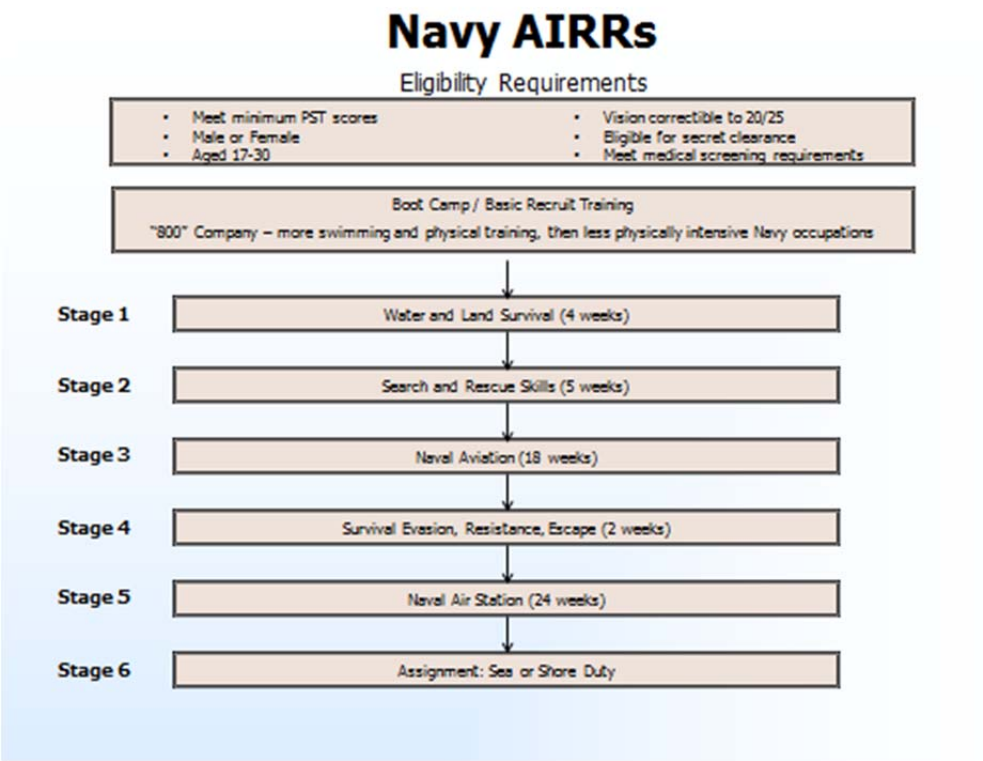


Figure B.3 Navy AIRR Screening Lifecycle



Enlisted Screening and Assignment

The enlisted occupational assignment and screening process is the same for EOD and Navy Diver (Navy Divers) recruits, and 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.

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 Navy diver 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 Navy Diver. 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 Navy Divers 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 street recruits. In a year with minimal street attrition there would be no fleet accessions. The Navy EOD and Navy diver commanders select which fleet members to send to training from among the 5-10 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 their 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-9 minute swims and runs, 20+ 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 males and females, were swim: 8:37; Push-ups: 103; Sit-ups: 95; Pull-ups: 16; Run: 9:33. Per the AIRR community manger, many male applicants complete their swim in 6:30, run in 8:00, 100 pushups and 10 pull ups.

Therefore, according to EOD and Navy Divers 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 Navy Divers community leaders indicated that even "if all the standards were increased by a 1/3, 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 Navy Divers, including the acceptance of street and fleet accessions, and inclusion of AIRR, EOD and Navy Divers recruits in the basic recruit training's 800 company. In particular, demonstration of the minimum PST standards is typically not sufficient to warrant a contract and invitation to enter training for AIRR recruits.

A difference between the EOD/Navy Divers 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, whether they have waivers, legal issues, region, recruiting district, program they want, age, height. The autoqual cutoff score is 1750, though we were told that waivers can be given even if that number is not reached. The history behind the autoqual as well as the actual formula is unknown to us.[23]

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

In recent years, the draft goal has included recruitment of 20 women annually, or just under 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.

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

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 are always very competitive and typically include 3-4 autoqual females. 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, for example 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 (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 females on the job, however 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 1-2 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.

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	NA

2015	20	9	NA

NOTE: NA indicates that not enough time had passed for all of the students to have had an opportunity to graduate. As a result, final numbers of graduates would not yet be known.

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/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 leaders. For EOD, the board typically selects 27-28 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 males and 12 females. The female officers all scored above the average male applicants, including on pull-ups (anecdotally, the females who were chosen each completed well over 15 pull-ups).

Training

Male street accessions to EOD, Navy Divers 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 do sailors entering less physically intensive occupational tracks. All EOD, Navy Divers and AIRR females are part of a separate female division equivalent to the 800 division though they complete PT with the males.

Further training for both fleet and street accessions is 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 lbs. 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 (120-150 lbs. 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 to either AWR (tactical) or AWS (non-tactical) A-School.[24] Selection into AWR or AWS depends on what the 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. While the two paths have similar physical demands, AWR tends to be more mentally demanding and AWS more utilitarian.

[24] Females enter Air Division School with the February and June cohort only.

Though some data provided to us in our interviews shows that PST times are useful predictors of training completion for these occupations,[25] 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 Navy Divers 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 in the past few years attrition from Rescue Swimmer School has been around 40-45 percent. 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, that they cannot accomplish the physical requirements in the pool with full equipment and that they are not comfortable in the water. There is very little attrition after Rescue Swimmer School, though there are still a few (around 3-4 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 high attrition rate, mostly because they did not really want to pursue the AIRR occupation.

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

According to Navy Divers and EOD community managers, the PST screening standards have not been significantly modified in over thirty 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 Navy Divers and EODs. However, those studies generally have not found strong support for using the PST elements to screen personnel for entry into these occupations (for examples, see Marcinik, Hyde & Taylor, 1994 and Hodgdon, Beckett, Sopchick, Prusaczyk, Gorforth, 1998). There were limitations to those studies, however, including that the trainee participants had already been screened on the PST and 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 however is reviewed and updated regularly. Naval Education and Training Command (NETC) procedures require a review of training once every 3 to 5 years through a process called the "human performance requirements review" (HPRR). Detailed instructions for how to conduct the review are documented in official Navy policy.

---

[25] 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 Recruit Training, only 11 percent (37 of 333) prep graduates or took longer than 1300 seconds (i.e., 21min and 40 seconds), compared to 34 percent (129 of 381) of those who dropped.

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, they 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. Since Navy Manpower Analysis Center sets the original occupational standards[26] they also play a role in the revision of the given training component. In addition, following NAVEDTRA 135, annually NETC conducts a formal review on every course including test item analysis, student critiques, attrition, etc. NAVEDTRA 135 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.

[26] NAVPERS 18068F volume 1 lists occupational standards and volume 2 lists navy occupational codes.

References

Air Force Enlisted Classification Directory (AFECD), The Official Guide to the Air Force Enlisted Classification Codes, April 30, 2013.

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

Army Times, *Army sets 160 seat for female Ranger School volunteers*, by Michelle Tan, Staff writer, December 5, 2014. Accessed on September 16, 2015: http://www.armytimes.com/story/military/pentagon/2014/12/05/women-ranger-school-students/19950227/

Chairman of the Joint Chiefs of Staff, Acting Under Secretary of Defense Personnel and Readiness, "Plan for Integration of Female Leaders and Soldiers Base on the Elimination of the 1944 Direct Ground Combat Definition and Assignment Rule (DGCAR)," memorandum to the Secretary of Defense, Washington, D.C., April 19, 2013.

DA PAM 611-21. 10-31B. MOS 31B-Military Police, CMF 31, (downloaded 2016, Last modified on Apr 15, 2015 9:59 AM) https://www.milsuite.mil/book/docs/DOC-145386, (not accessible to the public).

Dan Lamothe April 20, 2014 First Army Ranger School with women opens with 16 passing initial test, Washington Post. http://www.washingtonpost.com/news/checkpoint/wp/2015/04/20/first-army-ranger-school-with-women-opens-with-16-passing-initial-test/

Dawes, Robyn M., and Corrigan, Bernard. Linear models in decision making. *Psychological Bulletin*, Vol 81(2), Feb 1974, 95-106.

Dempsey, CJSC, Gen Martin E, "Women in the Service Implementation Plan," Info memorandum to the Secretary of Defense, Washington, D.C., January 9, 2013.

Department of Defense, Office of the Under Secretary of Defense 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*, February 2012.

Donley, Michael B., Secretary of the Air Force, "Air Force Implementation Plan for Integrating Women into Career Fields Engaged in Direct Combat," memorandum to the Secretary of Defense, Washington, DC, 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. Access at:
http://www.rand.org/pubs/technical_reports/TR744

Hardison, Chaitra M., Susan D. Hosek, and Chloe E. Bird, *Defining Physical Standards for Physically Demanding Jobs: A Review of Methods*, Santa Monica, Calif.: RAND Corporation, RR1340/1-OSD, 2015

Hodgdon, James A., M. B. Beckett, T. Sopchick, W. K. Prusaczyk, and H. W. Goforth, Jr., *Physical fitness requirements for explosive ordnance disposal divers*, No. NHRC-98-31, San Diego,Calif.: Naval Health Research Center, 1998

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.

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

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, Naval Medical Research Institute, Bethesda, Md.: 1993

NAVEDTRA 135C, *Navy School Management Manual*, (March 2010) Naval Education and Training Command

NAVPERS 18068F Volume I and Volume II, Navy Enlisted Occupational Standards, Manual of Navy Enlisted Manpower and Personnel Classifications and Occupational Standards Office of Personnel Management, *Delegated Examining Operations Handbook: Guide for Federal Agency Examining Offices*, 2012. Downloaded 9/17/15 at:
http://www.public.navy.mil/asnmra/corb/PEB/Documents/References/Rating%20MOS/Navy%20Rating%20Manual%20Jan%202012/18068F%20(Enlisted)Jan12.pdf

Office of Personnel Management, *Assessment & Selection: Other Assessment Methods: Physical Ability Tests*, 2015. Accessed at:
https://www.opm.gov/policy-data-oversight/assessment-and-selection/other-assessment-methods/physical-ability-tests/

Prusaczyk, W. K., J. W. Stuster, H. W. Goforth Jr, T. 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. K., J. W. Stuster, H. W. Goforth, Jr, M. B. Beckett, and J. 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

Secretary of the Navy, *Marine Corps Women in the Service Review Implementation Plan*, May 2, 2013

Sharp, Marilyn, *Development of Military Occupation-Specific Physical Employment Standards*. Unpublished IRB protocol dated February 19, 2014

Sharp, Marilyn, *Development of Military Occupation-Specific Physical Employment Standards: Study 3*. Unpublished IRB protocol dated May 28, 2014

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

Tversky, Amos, and Daniel Kahneman. Judgment under Uncertainty: Heuristics and Biases. *Science*, New Series, Vol. 185, No. 4157. (Sep. 27, 1974), pp. 1124-1131.

U. S. Army, Airborne and Ranger Training Brigade website, accessed on September 16, 2015: http://www.benning.army.mil/infantry/rtb/

U. S. Army, United States Army Ranger website, Accessed on September 16, 2015: http://www.army.mil/ranger/

U.S. Army, Goarmy.com, *RASP 1 & 2*, accessed on September 16, 2015: http://www.goarmy.com/ranger/training/rasp.html

U.S. Department of the Army, *Military Occupational Classification and Structure*, Army Pamphlet 611-21,Washington, D.C.: 2007

United States Marine Corps, Training and Readiness Manual group (TRMG), *Charter Terms of Reference*, NAVMC 3500.106, 2011

United States Special Operations Command, Office of the Commander, "U.S. Special Operations Command Implementation Plan for Elimination of Direct Combat Assignment Rule," memorandum to Chief of Staff, U.S. Army, Commandant, U.S. Marines Corps, Chief of Naval Operations, Chief of Staff, U.S. Air Force, Washington, D.C., March 22, 2013.

USASFC, Airborne website (a) accessed on September 16, 2015: http://www.soc.mil/USASFC/HQ.html

USASOC, 75th Ranger Regiment website (b) accessed on September 16, 2015: http://www.soc.mil/Rangers/75thRR.html

Vickers, Ross R., James A. Hogdon, 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.