Establishing Gender-Neutral Physical Standards for Ground Combat Occupations

Volume 1. A Review of Best-Practice Methods

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

RAND CORPORATION

For more information on this publication, visit www.rand.org/t/RR1340z1

Published by the RAND Corporation, Santa Monica, Calif.
© Copyright 2018 RAND Corporation
RAND® is a registered trademark.

Cover: U.S. Army image from Reuters.

Limited Print and Electronic Distribution Rights

This document and trademark(s) contained herein are protected by law. This representation of RAND intellectual property is provided for noncommercial use only. Unauthorized posting of this publication online is prohibited. Permission is given to duplicate this document for personal use only, as long as it is unaltered and complete. Permission is required from RAND to reproduce, or reuse in another form, any of its research documents for commercial use. For information on reprint and linking permissions, please visit www.rand.org/pubs/permissions.

The RAND Corporation is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest.

RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

Support RAND
Make a tax-deductible charitable contribution at
www.rand.org/giving/contribute

www.rand.org

Preface

Since the establishment of the all-volunteer force in 1973, representation of women in the U.S. 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that the last remaining policy restricting the service of women, the direct ground combat exclusion rule, would be rescinded.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 assignment for which they can meet the occupational standards. The SecDef also directed the military services to validate their occupational standards to ensure they appropriately reflect occupational requirements and are gender neutral.

This study 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 volume of the project report, identifies the best-practice methods for developing physical standards relevant to these military occupations. The second phase of the study, documented in the second volume, reviews the methods the services used to validate their occupational standards in response to the SecDef's guidance.

In December 2015, the Office of the Secretary of Defense released a preliminary version of this report on its website at the time the SecDef announced the decision to open previously closed combat occupations to women. This final version of the report incorporates minor changes resulting from RAND's quality assurance process and has been copy-edited and proofread for clarity and ease of reading.

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.

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

Contents

Preface iii
Figures ix
Tables ix
Summary xi
Acknowledgments xxi

CHAPTER ONE
Introduction 1
Why Standards? 5
Standards Versus Selection by Rank Ordering 9
Occupation-Specific Physical Standards and Entry into Military Service 11
Study Approach 13
Organization of This Report 14

CHAPTER TWO
Methodological Approaches to Establishing Physical Job Requirements 15
Six-Stage Process 17

CHAPTER THREE
Identify the Physical Demands of the Job 21
Methods for Conducting a Job Analysis 21
Reasons to Conduct a Careful Job Analysis 24

CHAPTER FOUR
Identify Potential Screening Tests 29
Taxonomies of Physical Aptitudes 30
Selecting Candidate Tests 33

CHAPTER FIVE
Validate and Select Tests 39
Construct Deficiency and Construct Irrelevance 41
Content Validity 44
Criterion-Related Validity 46
Convergent and Discriminant Validity 50
Fairness: Adverse Impact and Predictive Bias 52
Additional Considerations in Collecting Validation Evidence 55

CHAPTER SIX
Establish Minimum Scores 63
Use of Job Analysis Data to Set the Minimum Score 64
Use of Expert Panels to Set the Minimum Score 65

CHAPTER SEVEN
Implement Screening 75
When the Test Should Be Administered 77
Standardize the Test Administration Procedures 78
Informing Applicants About the Test 79
Consider Phasing the Test in Gradually 79

CHAPTER EIGHT
Confirm Tests Are Working as Intended 81
Institutionalize Research to Support Policy Changes 81
Conduct Additional Research as Needed 84
Ongoing Personnel Research Efforts Are Not New 85

CHAPTER NINE
Final Thoughts 87
Volume II of the RAND Study 90

Glossary 91
Abbreviations 97
References 99

Figures and Tables

Figures

S.1. Six Stages in Developing Physical Standards xiv
1.1. Enlistment and Initial Training of Military Enlisted Personnel 12
2.1. Six Stages in Developing Physical Standards 18
5.1. Conceptual Validation Linkages 40
5.2. Construct Deficiency 42
5.3. Construct Irrelevance 42
5.4. Illustration of a Hypothetical Selection Test Resulting in Underprediction of Female Performance 54

Tables

3.1. Example Considerations in Identifying the Physical Demands of the Job 26
4.1. Fleishman's Physical Ability Domains 31
4.2. Examples of Tests Used to Measure Physical Abilities in the Different Domains 33
4.3. Example Considerations in Identifying Potential Screening Tests 36
5.1. Example Considerations in Validating and Selecting Tests 56
6.1. Example Considerations in Establishing Minimum Test Scores 72

7.1. Example Considerations in Implementing Screening 76
8.1. Example Considerations for Confirming that Tests Are Working as Intended 82

Summary

On January 24, 2013, the Secretary of Defense (SecDef) 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). In announcing the decision to eliminate the rule, the SecDef 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.

Three years following that announcement, previously closed occupations were opened to women who can meet occupation-specific, gender-neutral standards of performance. In preparation for the opening of those positions, the Joint Chiefs of Staff established key requirements for implementing this policy change that had to be met prior to opening the occupations. These included validating performance stan-

dards for military occupations, with special attention to those occupations closed to women.

The Joint Chiefs of Staff highlighted two distinct but important principles to guide the services work. First, any existing gender-neutral physical standards would need to be validated. This was particularly relevant given that there were some existing gender-neutral physical standards in place in these occupations at the time of the announcement. However, these may not have been established using sound empirical evidence, and documentation of such evidence (if it existed) may not have been retained. In those cases, new evidence to support the existing standards would need to be collected and documented to address the call from the SecDef. The second requirement was that gender-neutral physical standards would need to be established where none had existed. Prior to the policy change, gender alone was serving as a proxy screener for physical ability in many jobs. However, with that proxy removed and the jobs open to both genders, the services would be faced with the possible inclusion of larger numbers of personnel who could not meet the physical demands. The services were given nearly three years to validate their standards and prepare to implement the policy change.

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 that can be tailored to the needs of the military. The second objective of the study was 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 used to make selection decisions, i.e., decisions made that can 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 was to ensure that their standards are 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. DoD can expect similar scrutiny as it embarks on the process of developing gender-neutral physical standards—so, for this reason, it wishes to employ appropriate methods in this endeavor. To assist the military services in developing general and occupation-specific standards relevant to performance, this report describes methods related to physical standard development.

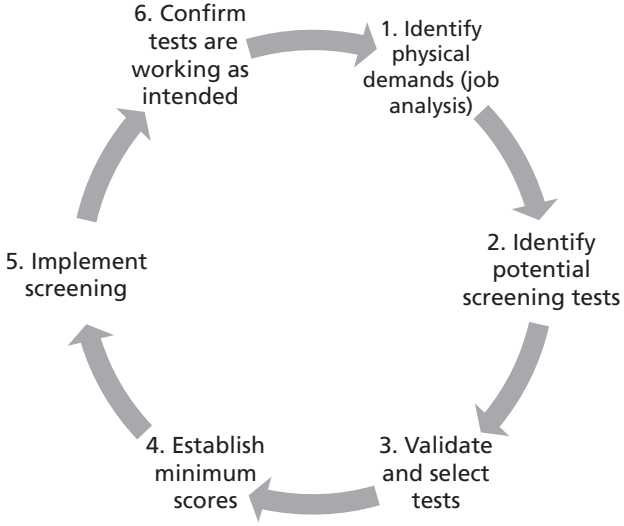
Methodological Approaches to Establishing Physical Job Requirements

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 organizational psychology, exercise physiology or a related field, psychometrics, and statistics. These experts rely on the knowledge of subject-matter experts, 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 in this report and, importantly, the documentation of the actions taken are critical to developing defensible 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

Figure S.1
Six Stages in Developing Physical Standards



RAND *RR1340/1-S.1*

of a job analysis serve as the foundation for nearly all types of human resource management activities, to include an organization's selection system. Job analyses can be conducted in several different ways. Some are worker-oriented approaches that focus on what workers do in performing their jobs; others are job-oriented approaches that focus on what workers accomplish in their jobs. Both approaches are valid and result in the collection of distinctly different types of information. Choosing among these alternatives, as well as determining how data are collected and what experts are called on to assist in the process, should be driven by the goals for the job analysis.

In establishing gender-neutral requirements for entry into physically demanding jobs, the focus is on applicant selection. Therefore, the goal of the job analysis should be to design an appropriate selection system. It should identify and describe in detail the physically demanding tasks the applicants would need to be able to perform in the job. In this context, task-level detail 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. If a job analysis is to be used for a different purpose (e.g., continuation in a job), it needs to be designed with that purpose in mind.

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

The next step in developing physical standards is identifying potential tests that might be used to screen job applicants (or job incumbents). In this context, we use screening to refer to evaluation of individuals' physical skills relevant for performing job tasks. Many factors weigh into this decision, but one important consideration is whether research and theoretical support exist for a test'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 of implementation 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 often relevant in choosing between high- and low-fidelity tests. All relevant costs must be considered, to include equipment costs, manpower costs, and validation costs. Feasibility relates to how difficult it is to implement a test in multiple locations. Cost and feasibility are of particular concern to the military services in, for example, considering whether to scale up an occupation-specific test for use by recruiters. Furthermore, because the military has many different physically demanding jobs, it faces unique challenges in selecting a set of tests for initial job classification. Using high-fidelity tests, in this context, may well be cost-prohibitive. Instead, administering a series of simple tests that can generalize across more than one job may be a more-feasible approach.

Where physical standards already exist for the occupation, the test(s) already being used should be included. To guard against the possibility that standards based on these tests prove not to be valid, other potential tests should 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 in a personnel selection context is to provide evidence that the selection test predicts important outcomes on the job.

Best practice requires evidence be accumulated to support claims that a test measures what it is intended to measure and that its scores can be used for selection. There are various types of validation evidence that an organization can collect, and each piece of evidence lends additional support to that claim. Validation evidence helps to answer several questions: Does the test fully capture the relevant characteristics of the physical requirements? Is there a clear relationship between test scores and outcome measures? Do the outcome measures capture

important job outcomes? If tests are deficient, then candidates 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 and methodological expertise and a careful design before data collection begins to ensure results are statistically sound (enough statistical power, representative of the population, etc.). In addition, bias against key subgroups (e.g., women or minorities) should be explored, which may require statistical oversampling of those subgroups. 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 reflect acceptable performance on the job. The goal in this step is to determine the minimum test scores that correspond to acceptable on-the-job performance. Test scores should be anchored to a concrete level of performance, such as lifting a certain number of pounds or running a specific distance within a certain amount of time. Minimum scores should be 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 contexts, 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 might 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 use the screening tool in personnel selection. But a number of key issues should be addressed during the implementation stage to ensure that the test is implemented in a manner consistent with the results of the validation and standard-setting efforts.

The timing of test administration can influence results. Tests administered far in advance of the predicted work should have 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 underpredict performance for everyone unless training effects are accurately taken into account—something that should be included in the validation process. It is also important to standardize test administration procedures so that each person has an equal opportunity to demonstrate his or her capability on the test regardless of where it is being administered. Key to standardization is creating clear documentation of the proper administration procedures and ensuring the equipment and testing environment are the same at all test locations.

Other important factors during implementation include informing applicants about the test so they have an equal opportunity to prepare. In addition, when new tests are instituted, an organization might 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 ongoing research to regularly confirm that tests are working as intended. Even the best research designs leave some questions unanswered. New, unanticipated questions can arise after implementation. Some studies are feasible only after a test has been implemented. Changing technology and missions can signifi-

cantly alter the requirements of the job. In addition, new research findings can emerge that suggest changes in testing policies. For all these reasons, the research effort should be treated as an ongoing process—one that continues long after a test has been implemented. Ideally, research efforts examining all stages of the standard-setting or validation process would be institutionalized as part of a regular operational data-collection activity for each occupation—a process not new to the military services.

Final Thoughts

The methods for establishing physical standards for specific occupations involve this six-stage process. The first four stages contribute to the initial development of the standards—the tests and minimum test scores that will be employed in selecting among applicants for entry into an occupation or among job incumbents for continuation in the job. 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 and men. When an occupation has been closed to women, the developers of standards must find a pool of women with related training and experience to represent women who might enter the occupation in the future.

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

Acknowledgments

We would like to thank Lt Col Mark Horner and Rennie Vasquez, our project officers during this first project phase 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 Beyler, 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 to make 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, Phoenix Voorhies, and Jason Way 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, Curt Gilroy, and Bryan Hallmark provided very helpful comments on an earlier draft.

CHAPTER ONE

Introduction

On January 24, 2013, the Secretary of Defense (SecDef) 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 1994 rule restricted the 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 that include physically demanding tasks the "vast majority" of women cannot do (SecDef, 1994).

Three years following this announcement, previously closed occupations were opened to women who can meet occupation-specific, gender-neutral standards of performance. Section 543 of the National Defense Authorization Act (NDAA) for 1994 mandated gender-neutral occupational standards to qualify individuals for any military occupation open to men and women and gender-neutral "specific physical requirements" for open occupations in which performance depends on "muscular strength and endurance and cardiovascular capacity." The fiscal year (FY) 2015 NDAA required that the "gender-neutral occupational standards being developed by the secretaries of the military departments (1) accurately predict performance of actual, regular, and recurring duties of a military occupation; and (2) are applied equitably to measure individual capabilities." These gender-neutral standards were to be developed, reviewed, and validated no later than September 2015, as specified in the FY 2014 NDAA (Section 524).

In preparation for the opening of ground combat positions to women, the SecDef and the Joint Chiefs of Staff established guiding principles for implementing the policy change; consistent with the congressional requirements in the 1994 NDAA, these included *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. Note that this direction highlighted two distinct but important principles to guide the services' work. The first was that any existing gender-neutral physical standards would need to be validated. This was particularly relevant given that there were some existing gender-neutral physical standards in place in these occupations at the time of the announcement. However, these may not have been established using sound empirical evidence, and documentation of such evidence (if it existed) may not have been retained. In those cases, new evidence to support the existing standards would need to be collected and documented to address the call from the SecDef. The second requirement was that gender-neutral physical standards would need to be established where none had existed. Prior to the policy change, gender alone was serving as a proxy screener for physical ability in many jobs. However, with that proxy removed and the jobs open to both genders, the services would be faced with the possible inclusion of much larger numbers of personnel who could not meet the physical demands. The services were given the following three years to validate existing standards and establish new ones. This was to occur prior to implementation of the policy change in 2016.

Military service is physically demanding, and the occupations that were 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 SecDef directed that the physical standards set for all military occupations be gender neutral. These 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 traditionally have 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.[2] 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. These occupation-specific standards are the focus of our study and must be gender neutral.

The challenge in establishing occupation-specific physical standards is to determine: (1) the physical capacities required to perform the job, (2) the most-suitable tests for assessing the relevant capacities,

---

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

[2] It could be reasonable to continue to use gender-specific standards for physical fitness requirements for enlistment and continuation in service, while still requiring that occupation-specific physical standards be gender neutral, if the goals of the two standards are different. In the case of occupation-specific standards, the goal is to ensure that people are capable of performing a specific set of tasks required of everyone or considered critical to performance on the job. In other words, a job-specific requirement means that it does not matter whether someone is male or female. If they can accomplish the requisite tasks, then they are qualified for the job; if they cannot, then they are not qualified for the job. Fitness standards, on the other hand, could be in place in the military for a variety of other reasons that have little to do with accomplishing a specific job requirement. For example, these could include a desire to maintain a culture of military discipline, bearing, and appearance; to keep health care costs to a minimum; to ensure personnel are not likely to be hampered by chronic illness; and to ensure that the personnel hired reflect the portion of the U.S. population who are at the peak of their health. All of these goals can be achieved using screening tools that evaluate someone's overall medical health and fitness. However, in general, medical health and fitness measures used for these purposes are gender-normed and age-normed. So, for example, the recommended resting heart rates for healthy men and women are different, and the recommended resting heart rate increases with age. Thus, if the military's intent with fitness standards is to determine whether someone is at the peak of health, resting heart-rate guidelines should be tailored to the age and gender of the person in question. Put simply, there could be a good justification for using gender-specific fitness standards, assuming that goals such as these have been articulated.

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, therefore, they also can 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 also can 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 excluding 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 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.[3] Some standards are valid for screening people for a particular job, and some are not. *Valid* standards are those that distinguish applicants who are likely to be able to perform 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 not. Defensible standards are those that have been developed according to best-practice methods that have been shown to be valid for all relevant subgroups,[4] 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*[5] intended to predict future performance on the job. In this study, we are most concerned with standards used to make *selection* decisions—i.e., decisions made that can 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 (Public Law 88-352; Public Law 102-166) to develop standards free of bias against *protected groups*.[6] That is, if the organization chooses a

---

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

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

[5] 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.

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

method of selecting employees that results in selection of different proportions of each protected group of applicants, it must show that the standards 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/ethnicity and gender bias) does not apply to the selection of military personnel;[7] 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 group that an organization wants to protect.[8] And the methods for examining adverse 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:

- *Economic costs from reduced performance on the job.* If personnel arrive on the job and cannot perform the work, a second person might have to pick up the slack. Or a task might 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.

---

7 However, DoD is subject to Title VII with respect to its civilian employment practices.

8 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).

- *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 from injuries to self or others.* Even if lives are not lost because of inadequate performance, individuals can become injured or conduct a task in such a way that other service members also become injured. Lost time on the job from injuries can affect unit readiness or mission performance.
- *Training costs lost because of 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 ensure their standards address that each of the above costs are the same regardless of gender—i.e., standards that are *gender neutral*.[9] 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

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

> 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.

In conducting this work, it should be expected that the services will discover a need for changes to their existing standards. This is common in the process of validating standards. As noted earlier, many existing standards in these closed occupations may not have been validated using the type of detailed empirical evidence described in this report. As a result, it would not be surprising if new information was uncovered leading to improvements in the screening processes. In addition, it would not be surprising for entirely new screening criteria to be introduced as a result of the services data collection efforts, given that a whole new population of individuals with an entirely different physical ability distribution is being introduced into the applicant pool. The intent of the process is not to lower standards to accommodate this new population, but rather to ensure that the integrity of the performance of those who make it into the occupation is maintained or even improved. Secretary Panetta spoke directly to this point in his comments, as noted earlier in this chapter, when he stated that the goal is not to reduce the qualifications for the job, but rather to ensure that those in the job are the most capable.

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 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,[10] and standards may be established for determining who will be allowed to continue in the occupation later in their careers. 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 gender-neutral physical standards 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 open to women for decades, were just recently opened, or may be opened in the near future.

Standards Versus Selection by Rank Ordering

It is worth noting that in some jobs, the goal is not to establish a minimum standard, but rather to take the best candidate(s) among those who have applied for the job until the number of positions has been filled. That process is known as top-down selection. In top-down selection, evidence supporting a robust linear relationship showing that higher scores on the screening tests used are associated with higher levels of performance on the job is necessary to defend the use of top-down selection. However, in many cases, top-down selection may not be preferred; in some cases, it may not be defensible. Instead, establishing minimum scores for selection into a job may be a better alternative.

A variety of circumstances should lead an organization to consider establishing minimum score standards instead of or in addition to top-down selection. First, some jobs have a minimum performance require-

[10] 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.

ment below which selection should stop, regardless of the number of openings available. Jobs in which performance below a certain level could pose a danger to others is one example of a circumstance where drawing a minimum standard would be critical. In other jobs, the organization also may need to screen for performance qualities other than physical ability (such as persistence, cognitive ability, conscientiousness, or technical skills), as long as a minimum threshold of physical performance can be met. In these cases, establishment of minimum scores could also be relevant. Moreover, in some jobs there may be no real benefit to gains in performance beyond a certain threshold, especially at the very high end of the score distribution. For example, if a job requires lifting a particular weight regularly (but nothing of a higher weight), once someone has passed the threshold of being able to lift that weight with ease, there is no further gain to the job when someone is even stronger. Jobs fitting these circumstances are especially good candidates for establishment of minimum physical ability scores, given the potential adverse impact of the screening against women. When there is no evidence to show that the relationship between physical ability test scores and performance is linear along the entire score distribution (particularly above and around the point at which top-down selection would yield selection/rejection decisions), establishing minimum scores may be more defensible.[11] Lastly, in some jobs, training is a hurdle that must be passed in order to be qualified, so the training performance standards should remain constant from class to class to ensure fairness and consistency in the quality of personnel produced by the training. Furthermore, minimum standards for whom is allowed to even attempt training also may need to be established to ensure a minimum level of class quality. In these cases, establishing minimum training standards required for graduation from training would be relevant, as would establishing minimum training entry standards.

Thus, whether top-down selection or minimum standards are needed is partly dependent on the goal of the selection process and the needs of the job. This should be clearly articulated in advance of con-

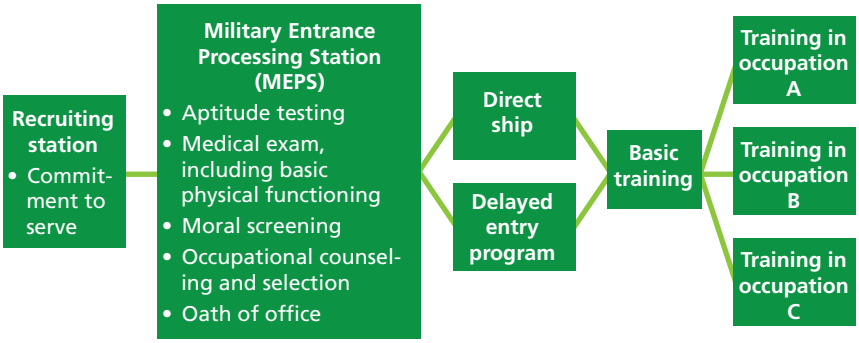
[11] See the *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology, 2003).

ducting a validation effort, and any time top-down selection is chosen over establishing minimum standards, a clear rationale for why it is appropriate should be provided, along with evidence that supports that rationale. In the military, minimum standards are established for many reasons, including to maintain safety, ensure minimum levels of performance are maintained in an occupation, ensure consistency in training graduation requirements from class to class, and to maintain a minimum level of trainee quality. In cases where minimums are relevant, the entire methodology described in this report is relevant. In cases where top-down selection is well justified, researchers could skip the process of setting standards, but all of the other methods for validating selection processes described in this report would still apply. The remainder of this report assumes that setting standards is the goal and addresses that explicitly; however, we do note the point at which the recommended methodology would diverge if top-down selection alone is being pursued.

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 specify at which point in the entry process the screening will be done and standards for entry 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.

Figure 1.1
Enlistment and Initial Training of Military Enlisted Personnel



SOURCE: Authors' analysis.
RAND *RR1340/1-1.1*

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 when 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 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.

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 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 addressed two research objectives. First, in this volume of the project report, we describe best-practice methodologies for establishing standards for physically demanding jobs that can be tailored to address the needs of the military. No single method applies to all jobs; instead, the method would need to be tailored to the job. So this report was intended to provide the services with insights regarding the variety of appropriate methods, while still leaving them free to choose methodologies uniquely tailored for the occupation. Second, using the framework in this report as a guide, we reviewed the methodologies used by the mili-

tary 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. The second volume of the report describes the methods and results of this review.

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.
- Chapter Nine summarizes the key steps in the standards development process and describes next steps in conducting our study.
- A glossary of the terms in this report can be found in the Appendix.

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 (also known as personnel psychology[1]) 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 [CFR], 1978c).[2] The Uniform Guidelines do not apply to the military, but they can 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:

---

[1] Personnel psychology is an area of study within industrial and organizational psychology.

[2] 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" (Title 29 of the CFR, Part 1607, 1978c; Title 41 of the CFR, Parts 60-63, 1978d; Title 28 of the CFR, Section 50.14, 1978b; Title 5 of the CFR, Section 300.103, 1978a).

- *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, 2014). 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 physical skills and abilities. This applicability is noted explicitly in the *Principles*. The methodological approaches we describe are consistent with those advocated in both the *Principles* and the *Standards*.[3]

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,

[3] 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.

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 importantly, it can provide insights into the types of tests that might be useful for employment screening and selection.

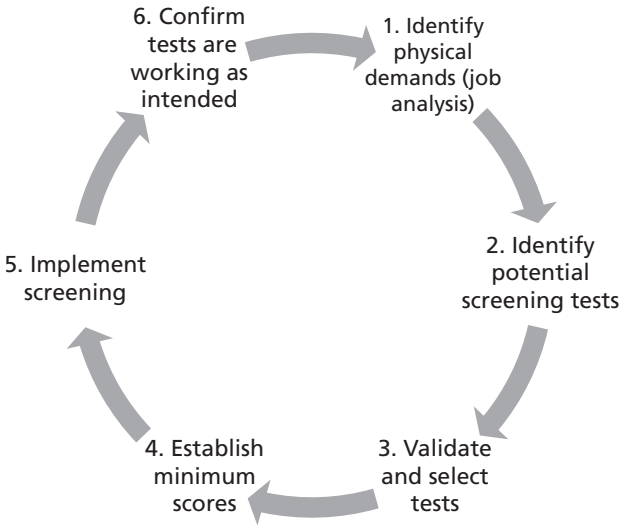
Whereas the field of personnel selection offers the methodological approach, the field of physiology serves as the starting point for much of the content applied in the methodology.

Six-Stage Process

We organize the overall approach for developing physical standards into six general stages, as depicted in Figure 2.1 and described as follows. These steps also provide a useful framework for evaluating any standards already in place. For those standards, the services should review their existing evidence in support of each of these steps and consider supplementing past efforts if any gaps 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. 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 exist, stage 2 may not 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 and 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., performance, injury rates, productivity). From the results of validation studies, identify the

Figure 2.1
Six Stages in Developing Physical Standards



RAND *RR1340/1-2.1*

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 valid for all groups. 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. This step is necessary when an organization aims to set absolute standards for ensuring that a specific level of performance on the job is maintained, or when standards are used in addition to or in place of top-down selection to screen personnel.
- **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 the tests are working as intended. Verify whether test administration in practice adheres to the established guidelines. 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.

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 for 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 in a job is known as *job analysis*.[1] 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 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.[2] The best-known methods are:

---

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

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

- 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 these tasks on such factors as importance and frequency.
- 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 approach used to develop DOL's Occupational Information Network (O*NET) database of occupational requirements and worker attributes.[3]
- 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 variety of topics, including the work environment, the types of information sources used on the job, mental processes, and work output. The result can be highly detailed, but it does not provide task descriptions 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 Analy-

[3] Available at the O*NET OnLine website (National Center for O*NET Development, undated).

sis 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 also can 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 questionnaires. The people who serve as experts also vary. 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 new jobs, supervisors or instructors 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 necessary level of detail should instead be driven by the goals for the use of the results.[4] However, it is important to note that job analysis data adequate for one activity may not be acceptable 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 selection system for physically demanding jobs. Therefore, the primary goal of the job analysis should be to identify and describe in detail the physically demanding tasks the applicants need to perform on the job. Task-level detail specific to each occupation is ideal for a sound defense of a selec-

[4] 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.

tion system.[5] 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 can 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, can 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

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

> 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)

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.

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 in Table 3.1 would likely be useful for designing training simulations, 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 what was originally intended.

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 important tasks performed on the job (not just those that are physically demanding). Assess the importance and frequency of each task 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.
Are some physically demanding tasks considered critical to successful performance, even if infrequent?	Assess the importance/criticality of the tasks. Include those considered critical functions of the job, even if they occur infrequently.
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 type 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 drill in an armored vehicle). 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.

Table 3.1—Continued

Considerations	Potential Resolution
Are individuals who are selected as experts for the job analysis sufficiently knowledgeable? Is the number of experts too low 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-level versus lower-level positions, or alternative ways to accomplish the same task?	Include participants from a 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 job levels to determine whether meaningful differences exist. 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, or experienced in performing the types of tasks that occur in that occupation) to observe and learn about the job and include their perspectives in the job analysis.

SOURCE: Authors' analysis.

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 indicated in Chapter Two, even where physical selection standards exist, 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 area, 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.[1] In cases in which an employer chooses to devise a new test, one for which research does not 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.

[1] 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.

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 facing 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.[2] The concept underlying Fleishman's research is that people can score high on one physical aptitude without necessarily scoring high on 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.

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

---

2 For a review, see Myers, Gebhardt, Crump, and Fleishman, 1993.

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

Table 4.1—Continued

General Area	Basic Ability or Domain	Domain Definition
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.

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	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

NOTE: The Illinois agility test measures the time it takes to run around a weaving course, and the 505 agility test measures the time it takes to run five meters, turn 180 degrees, and run back five meters.

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 can-

didates to perform a variety of typical firefighting tasks (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*[3] 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 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

[3] 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.

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.

Feasibility is a third consideration, one 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 laps may not be comparable to those 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 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 for validation. 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.

Table 4.3—Continued

Consideration	Potential Resolution
Do other tests have lower adverse impact on women?[a]	Include tests that could be useful predictors but are known to have smaller race/ethnicity or gender differences. Compare validation results with those that have larger race/ethnicity or gender differences.

SOURCE: Authors' analysis.

[a] For more on discussion on adverse impact, see the section titled, "Fairness: Adverse Impact and Predictive Bias" in Chapter Five.

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 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).[1] The three key types of validation evidence are

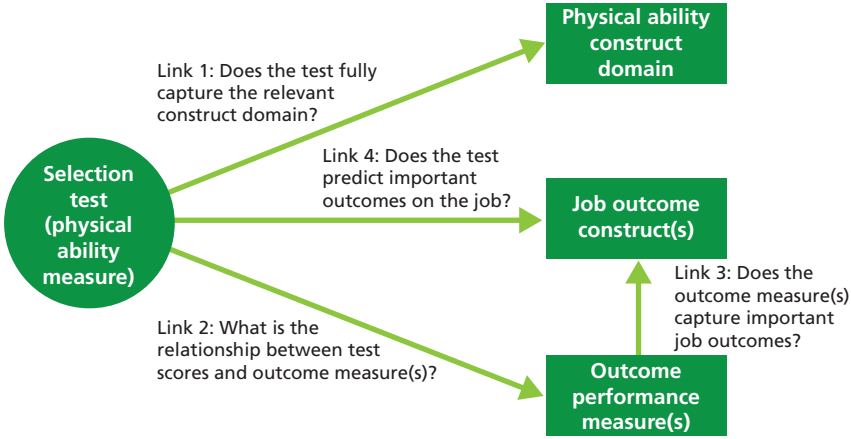
- evidence of *content validity* is evidence that a test covers the job content domain of interest
- evidence of *criterion-related validity* is evidence that the test predicts important organizational outcomes
- evidence of *convergent* or *discriminant validity* is evidence that a test measures what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

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

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.[2] 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 test predicts important outcomes on the job). No single method of validation can provide complete support for that link. Instead, amassing

Figure 5.1
Conceptual Validation Linkages



SOURCE: Adapted from Binning and Barrett, 1989.
RAND *RR1340/1-5.1*

---

[2] For example, there would be strong theoretical support for the use of a test 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 amassed, the stronger the support.

information that 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 further later in this chapter.

Construct Deficiency and Construct Irrelevance

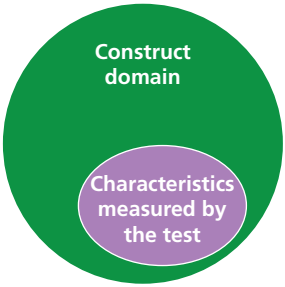
The first step in any validation effort is to clearly define the *constructs* (i.e., the concepts or characteristics) one intends to measure. Verbal and mathematical aptitude, personality, job performance, finger dexterity, and physical strength are examples of broad constructs that have been explored in personnel research. However, validation 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—i.e., whether link 1 in Figure 5.1 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.[3] 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 measure strength that measures upper-body strength but not core or lower-body strength would also be construct deficient. 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 also tap lower-body strength (if it did, that part of the test would be construct irrelevant; see Figure 5.3).

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 who are not capable of performing on the job (i.e., *false positives*) may be

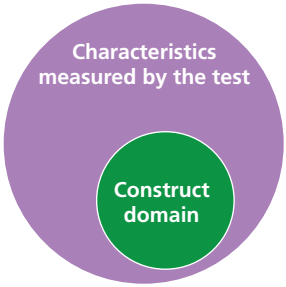
[3] See Messick, 1989, for more on these concepts.

Figure 5.2
Construct Deficiency



SOURCE: Authors' analysis.
RAND *RR1340/1-5.2*

Figure 5.3
Construct Irrelevance



SOURCE: Authors' analysis.
RAND *RR1340/1-5.3*

selected or candidates who would have been capable of performing the job (*false negatives*) may be rejected in higher numbers than otherwise might be the case.

Construct irrelevance—or *construct contamination*, the measurement of something other than what was intended—is also problematic (see Figure 5.3). It also 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 it measured a specific construct domain but, during administration in a real testing environment, it actually measures motivation to perform, then 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 is in reality. 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 overestimate 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 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 also could 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 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[4] and the extent to which it may be affected by construct irrelevance.[5] In making their judgments, SMEs should be instructed to consider the manner in which the test is administered. The following are examples of questions the SMEs should use: Does the test measure all aspects of the ability it is stated

---

[4] 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.

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

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?

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 it 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 the ability to perform the task is a necessary job condition 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 can differ in the extent to which they already have 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 or *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, retention, and 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.[6] 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., injuries, job performance, attendance) on job incumbents and administering the selection tests to those same incumbents. Concurrent validation evidence is not ideal. Because 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 might perform poorly on the test as applicants but could 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

---

[6] 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., graduation from a six-week training program).

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 or a related test already has 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 selected group (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.[7] 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 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).

---

[7] 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.

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, and 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; and disability rates)
- job performance (e.g., supervisory or peer ratings)
- 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; and 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 also may be justified. For example, in the case of physically demanding jobs, training dollars lost to attrition, medical costs, or time lost because of injuries could be argued as important organizational outcomes to be predicted from a selection test.

Nevertheless, measures that might superficially appear justified could 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.[8]

---

[8] Note that if a physical ability test was found to be an impure measure of a physical ability domain but instead was contaminated by some other construct domain, such as intelligence

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 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 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 and lower-body strength.[9]

---

or motivation, it could turn out to be a better predictor of performance on the job because of the added contamination. However, the effect of this or any other type of contamination on a test's predictive validity would need to be examined empirically to determine its effects to include whether it adds predictive bias. Some forms of contamination can reduce validity, and others can enhance it. Similarly, some can increase adverse impact, and some can reduce it. Regardless, irrelevant variance can change conclusions regarding the validity of test scores and, when identified, it warrants closer examination.

[9] If the relationship is calculated by gender, convergent-validity and discriminant-validity estimates might be expected to differ in this example. 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.

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, 2014), although it is most often applied to further theoretical understanding for the constructs measured by the test. However, such theoretical understandings for the tests can be important in defending the use of an employment test that might not stand up to criticism when 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 One, tests used for occupational screening should be fair. There are multiple dimensions to fairness (Joint Committee on Standards for Educational and Psychological Testing, 2014), but in practice the issue "is typically addressed by statistically examining test results for evidence of bias. 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* [emphasis added] by subgroup. If the predictions are equivalent (i.e., no differences in [estimated relationships between test scores and performance measures]), then there is no bias" (Jeanneret and Zedeck, 2013, pp. 602–603). *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.[10] For example, if 70 percent of male appli-

[10] 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-per-*

cants 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, 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*.[11] 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.[12] For example, if, for men, a score of 10 on a strength test suggests that they will fail, and the same test is used

---

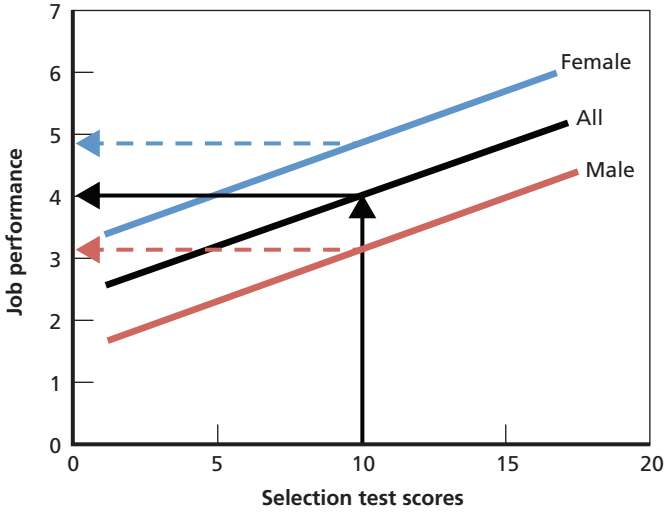
cent rule. The 80-percent rule does not apply to the selection of military personnel; however, similar principles regarding adverse impact still are applicable given that equal opportunity is strongly supported in the military.

[11] 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, 2014).

[12] 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 might lie in the choice of the predictor construct.

for both men and women, then a 10 for women should have the same expected outcome—namely, failure. However, if a study shows that a score of 10 would predict that women on average would succeed on the job when men with the same score on average would fail, the test would be underpredicting female performance (see Figure 5.4 for an illustration). In circumstances where underprediction for women is occurring, setting the same minimum score for men and women would screen out more women than men who would be able to do the job. 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

Figure 5.4
Illustration of a Hypothetical Selection Test Resulting in Underprediction of Female Performance



SOURCE: Authors' analysis.
RAND *RR1340/1-5.4*

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, 2014) for more information about how to conduct SME panels for evaluating bias.

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.[13] 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 per-

[13] 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 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.

Table 5.1—Continued

Consideration	Potential Resolution
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.
Can scores on the test be used for top-down selection or are they better suited to use as a criterion-referenced screening standard?	If the test is to be used for top-down selection, evidence supporting the conclusion that higher scores on the test result in better performance on the job is needed for the part of the test performance distribution at which selection using top-down selection typically occurs. If the selection process usually results in only the top 10 scorers out of 100 candidates being selected, for example, evidence showing that the top 10 are going to be more successful on the job than the top 20 or top 30 will be critical to supporting the decision to take just the top 10. Alternatively, if establishing a concrete floor for the minimum level of performance on the test is the goal, approaches for setting standards should be pursued. These are described in the next chapter.

SOURCE: Authors' analysis.

formance. 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 might show that some demographic populations (such as women or other groups) should be over-

sampled. In cases in which groups are oversampled, complex sampling statistics need to be applied in the subsequent analyses.

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.[14] 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) or 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

[14] 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.

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.

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.

Establish Rationale and Evidence to Justify a Test's Use for Top-Down Selection, If Relevant

As noted in Chapter One, there are essentially two ways that physical tests can be used to select personnel: for top-down selection (where candidates are selected in order of their scores—either alone or in combination with other information—until all available spots have been filled); and establishing minimum standards (where all those who are likely to meet a required performance level on the job are considered

qualified for entry, regardless of the number of positions to be filled). In many military jobs, both types of uses could be justified; however, much of the focus of the discussion surrounding the opening of combat jobs to women has been on establishing minimum job performance, occupational entry, or other screening standards. For justifying the use of tests for that purpose, we refer readers to the following chapter for the next step in the process of amassing evidence and a rationale for that use.

Nevertheless, some tests may be intended to serve as top-down selection tools only, or may be intended to serve both purposes (for top-down selection and establishing a minimum floor for those considered qualified). In those cases, validation evidence showing that the test supports the contention that "more is better at all ranges of the distribution" is important, especially in the range of performance where the test is having an impact on selection. In other words, evidence showing that performance on the job gets better and better (i.e., it is increasingly valuable to the success of the organization) with more and more of the underlying skill or ability being assessed by the test at all ranges of the job performance continuum is one critical component of that evidence. For example, if lifting objects weighing 60 pounds all day is the requirement, but being able to lift more is not necessary or beneficial, then once people reach the threshold of being able to lift the 60 pound objects all day, there would be a leveling off of the gains in work performance that would result from increases in strength. Thus, a more-is-better argument at all levels of the strength continuum would not be supported.

Another critical component required to support top-down selection is evidence showing that the test is sufficiently accurate at all levels of the score range to distinguish differences in underlying skill or ability that relates to the increases in performance. For example, if the test does a good job of distinguishing performance of people at the bottom end of the distribution, but at the higher end it becomes much less reliable, it should not be used to distinguish among the top performers. This can be observed by exploring whether the relationship is linear across the entire score range (or positive and strong enough to justify the validity of the test for distinguishing performance levels at the rel-

evant score ranges) and whether the error associated with test scores is constant across the entire range of the distribution. If a smoothed regression line does not show the same consistent positive relationship at the upper end of the distribution, the test should not be used to distinguish among candidates at that point in the score continuum. In cases where the scores are not linear, or data are not available to support an argument of linearity across the entire range of scores, the test might be better suited for use in establishing minimums, as discussed in the following chapters.

Thus, if performance above a certain level has diminishing returns for an organization, or if the test fails to accurately distinguish who at a given level will be likely to be better on the job, then its use in top-down selection may not be justified. The validation stage is the point at which these issues should be explored, a rationale for the test's use for top-down selection should be articulated, and evidence supporting that use should be documented.

CHAPTER SIX

Establish Minimum Scores

Once a test or series of tests has been selected and a need to establish minimum standards (instead of or in addition to using the test for top-down selection) has been articulated,[1] the next step in the process is to establish the minimum scores that will reflect acceptable performance on the job.[2] 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 SecDef's emphasis on not "reducing the qualifications for the job" is important for determining what minimum 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

[1] For a discussion of when standards may be relevant instead of or in addition to top-down selection, see Chapter One.

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

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 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.

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 shows 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 consensus among SMEs that dragging a body 50 feet is considered an important part of a firefighter's job, and the test involves a simulation of dragging a 150-pound dummy 50 feet (and any other critical task characteristics that effectively simulate the conditions under which the activity is conducted), then theoretically those who accomplish the task pass. However, test minimums are sometimes not so straightforward.

Use of Expert Panels to Set the Minimum Score

Although this previous example 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 suspect may be an important part of the job, so one of the 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.

Standard-setting will require information other than the job analysis when the tests do not rely on simulated job tasks, there is no obvious overlap between the test and the contents of the job, or there is no obvious line distinguishing success from failure on the job. 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 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.

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 earlier), 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. How-

ever, 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[3] 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 prescribed as best practice for conducting a standard-setting study. 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. Some of those elements are:[4]

---

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

[4] 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.

- **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 race/ethnicities and genders. How many would be considered enough is largely dependent on the context and number of stakeholders to be represented. For example, locations can 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 earlier in this chapter, 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 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 in the study using an entirely different 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 its difficulty. The purpose of the test also should 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 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.[5]

- **Ask SMEs for feedback on the standard-setting process and the resulting minimums.** Part of the goal of standard-setting is to establish minimum test scores 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.

---

[5] 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]	Choose 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.

Table 6.1—Continued

Consideration	Potential Resolution
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 with 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 does not meet the minimums, and explore why they are still on the job. Is the test not sufficiently correlated with performance? 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.

SOURCE: Authors' analysis.

[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

- 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 administered in a manner 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 7.1.

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 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 given the same way. Train people in the administration procedures. Conduct regular checks to confirm adherence to those procedures. Fix any inconsistencies in administration identified during the checks.
Do all people know about the test? Those who know about it might 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.

Table 7.1—Continued

Consideration	Potential Resolution
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.
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 men, 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.

SOURCE: Authors' analysis.

When the Test Should Be Administered

The timing of test administration is important. Tests administered far in advance of the predicted work should have evidence to show that the time gap does not change the validity of the test or the interpretation of scores. For example, in some cases, scores collected before the time gap can 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 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, the amount of improvement is likely to be test-dependent and 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 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.[1] Standardizing test procedures across administrations and locations is one way to ensure fairness.

The first steps to standardization are creating clear documentation of the proper administration procedures and ensuring that the equipment and testing environment are the same at all locations. The procedures, equipment, and testing environments 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 adhering 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 verify they are following

[1] See the *Standards* (Joint Committee on Standards for Educational and Psychological Testing, 2014) for more information.

the procedures correctly. Providing explanations about the importance of adherence to the procedures is important and can help ensure that administrators conform to them. Some of the justifications could be 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.

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.

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, 2014, 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 types of information that should be provided include

- 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
- 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 in the test so that applicants have enough time to become familiar with it and prepare. 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.

CHAPTER EIGHT

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 and unanticipated questions may arise after implementation. Some studies are feasible only after a test has been implemented. Constantly changing technology and missions 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 also should 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.

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.

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.

SOURCE: Authors' analysis.

Reexamine Job Analyses

Job analyses should be redone on a regular basis to ensure that job requirements have not changed. For jobs 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 specified 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 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. 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. For example, pull-ups could be a good predictor of box-lifting capability on the job initially. This could occur because people who can do many pull-ups also tend to be generally physically fit and active in other ways. They likely spend time working out, focusing on many different upper-body muscle groups in addition to those needed to perform pull-ups. As a result, pull-up performance would be highly correlated with arm strength in general, which could make it a useful test for box-lifting, at least initially. However, if test-takers instead focus their training on doing well on the test and developing the capacity to perform pull-ups, it is possible that they could develop the specialized muscles to perform the required pull-ups but still struggle with box-lifting as they have not developed the rest of the needed upper-body strength involved in lifting boxes. If so, their pull-up scores would no longer be a good predictor of their expected box-lifting performance. If many people begin to train just that narrow muscle group to pass the test, the validity of the test in practice could go down over time. Therefore, information on whether individuals are training for the test and whether the type and amount of training affect the predictive validity of their scores should be collected. Such information should be used to determine whether 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 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. 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 also include elements for addressing other personnel issues.

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. 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., 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 sex 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)—i.e., 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 and 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 established guidelines. 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 when the testing is done improperly and as the occupational tasks and equipment change over time. Similarly, ensuring ade-

quate performance requires establishing appropriate physical standards for job incumbents based on the tasks they are expected to carry out over their careers.

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

- 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.

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 environments, and on appropriate test subjects drawn from the population of applicants, trainees, and job incumbents.

Volume 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 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

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 against members of certain groups because that characteristic is materially important to job performance. 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 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: Universal standards, regardless of gender, that are equally valid in predicting job performance in 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 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 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 c*riterion-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 a test or measure. See also *content validity, convergent and discriminant validity, criterion-related validity*.

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

work analysis: See *job analysis*.

Abbreviations

CFR	Code of Federal Regulations
CODAP	Comprehensive Occupational Data Analysis Program
DoD	U.S. Department of Defense
DOL	U.S. Department of Labor
FY	fiscal year
MEPS	Military Entrance Processing Station
NDAA	National Defense Authorization Act
O*NET	Occupational Information Network
OUSD(P&R)	Office of the Under Secretary of Defense for Personnel and Readiness
SecDef	Secretary of Defense
SME	subject-matter expert
VO_2 max	maximum volume of oxygen used during incremental exercise

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.imta.info/PastConferences/Presentations.aspx

Brannick, Michael T., Edward L. Levine, and Frederick P. Morgeson, J*ob 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, 1978a. As of March 12, 2013: http://www.gpo.gov/fdsys/pkg/CFR-2012-title5-vol1/pdf/CFR-2012-title5-vol1-sec300-103.pdf

———, Title 28, Judicial Administration, Vol. 2, Chapter I, U.S. Department of Justice, Part 50, Statements of Policy, Section 50.14, Guidelines on Employee Selection Procedures, 1978b. As of March 12, 2013: http://www.gpo.gov/fdsys/pkg/CFR-2010-title28-vol2/pdf/CFR-2010-title28-vol2-sec50-14.pdf

———, Title 29, Labor, Part 1607, Uniform Guidelines on Employee Selection Procedures, 1978c. As of March 12, 2013: https://www.gpo.gov/fdsys/pkg/CFR-2011-title29-vol4/xml/CFR-2011-title29-vol4-part1607.xml

———, Title 41, Public Contracts and Property Management, Parts 60-63, Uniform Guidelines on Employee Selection Procedures, 1978d. As of April 29, 2016: http://www.ecfr.gov/cgi-bin/text-idx?c=ecfr&sid=3b71cb5b215c393fe910604d33c9fed1&rgn=div5&view=text&node=41:1.2.3.1.3&idno=41

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*, New York: 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

Jeanneret, P. Richard, and Sheldon Zedeck, "Professional Guidelines/Standards," in James L. Farr and Nancy T. Tippins, eds., *Handbook of Employee Selection*, New York: Routledge, 2013, pp. 593–623.

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

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.

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.

National Center for O*NET Development, "O*NET OnLine," homepage, U.S. Department of Labor, undated. As of April 28, 2016: http://online.onetcenter.org

Panetta, Leon E., "Statement on Women in Service," speech delivered by Secretary of Defense Leon E. Panetta, Pentagon Press Briefing Room, Washington, D.C., 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.

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 to the Secretary of the Army, Secretary of the Navy, Secretary of the Air Force, chair of the Joint Chiefs of Staff, Assistant Secretary of Defense for Personnel and Readiness, and Assistant Secretary of Defense for Reserve Affairs, Washington, D.C., January 13, 1994.

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, P*rinciples 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://archive.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. 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