Draft Research Report

Defining Physical Standards for Physically Demanding Jobs

A Review of Methods

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

RAND National Defense Research Institute

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

Preface

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that the last remaining policy restricting the service of women, the direct ground combat exclusion rule, would be rescinded.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

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

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

Contents

Figures

Tables

Summary

On January 24, 2013, the Secretary of Defense and Chairman of the Joint Chiefs of Staff announced rescission of the 1994 Direct Ground Combat Definition and Assignment Rule (SecDef, 1994) and the intention to "integrate women into occupational fields to the maximum extent possible" (U.S. Department of Defense, 2013). The rule restricted assignments of women to occupational specialties or positions in or collocated with direct ground combat units below the brigade level, in long-range reconnaissance and special operations forces, and in positions including physically demanding tasks the "vast majority" of women cannot do (SecDef, 1994). In announcing the decision to eliminate the rule, the Secretary stated:

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

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

The Office of the Under Secretary of Defense for Personnel and Readiness asked RAND to help it understand how to evaluate job-specific physical requirements and establish gender-neutral standards for physically demanding jobs. Our study addresses two research objectives. The first is to describe best-practice methodologies for establishing gender-neutral standards for physically demanding jobs, tailored to address the needs of the military. The second objective of the study will be to review the methodologies being used by the military services to set gender-neutral standards. This report provides the results of work conducted toward the first research objective.

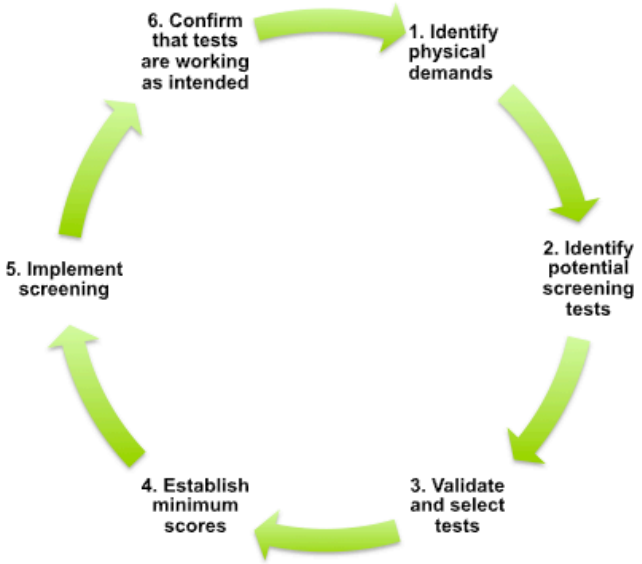
Throughout this report, we use the term *standards* or *physical standards* to refer to occupation-specific criteria that applicants must meet to enter or remain in a particular career field or specialty. We are concerned with standards that are used to make selection decisions—that is, decisions made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

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

Methodological Approaches to Establishing Physical Job Requirements

The methods for developing physical standards can be organized in six general stages (see Figure S.1). Each element in the process for establishing physical job requirements provides support for the use or exclusion of a set of selection procedures. Carrying out this process requires expertise in a variety of domains, including industrial and organization psychology, exercise physiology or a related field, psychometrics, and statistics. These experts rely on the expertise of subject matter experts from the occupation, who must be carefully selected to cover all types of work and work environments, and on appropriate test subjects drawn from the population of applicants, trainees, and job incumbents. The deliberate steps described here and, importantly, the documentation of the actions taken are critical to developing defensible physical standards.

Figure S.1. Six Stages in Developing Physical Standards



1. Identify Physical Demands

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

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

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

2. Identify Potential Screening Tests

Identifying potential tests that might be used to screen job applicants (or job incumbents) is the next step in developing physical standards. In this context, we use screening to refer to evaluation of individuals' physical skills relevant for performing job tasks. Many factors weigh into this decision, but one important consideration is whether research and theoretical support exist for a tool's use in a similar employment context. Test developers and employers should be aware of relevant research results—whether new tests are being explored or well-established tests are being considered.

Selecting the right tests in an employment context requires careful attention to which physical abilities are and are not required by the job. Once these are determined, a variety of factors come into play when selecting a test: fidelity to the job, cost, and feasibility are three of

the most important. Fidelity to the job refers to the similarity between the test and job tasks. High-fidelity tests have obvious overlap with the job and are often viewed as more fair by test takers. Low-fidelity tests have little observable similarity to job tasks but instead measure general physical abilities that may be relied on to perform job tasks. There can be some overlap in the two types of tests, and either type or a combination of both can be used effectively to screen job applicants.

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

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

3. Validate and Select Tests

The third step in developing physical standards is to validate potential tests and identify those with the highest validity and least adverse impact. In the personnel selection context, the term validate has a precise meaning. It refers to the act of accumulating multiple sources of research-based evidence to support a test's use for a particular purpose. The ultimate goal of validation is to provide evidence that the selection test predicts important outcomes on the job.

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

Collecting validation evidence is a complex process. When undertaking validation studies, an organization must document all aspects of the research study design and its results. These studies typically require considerable statistical expertise and require a careful design before data collection begins to ensure results are as accurate as possible and avoid bias toward any group of applicants. Finally, organizations should seek multiple sources of validation evidence whenever possible.

4. Establish Minimum Scores

The next step in the process is to establish the minimum scores that will reflect acceptable performance on the job. The goal in this step is to determine the minimum test score(s) that corresponds to acceptable on-the-job performance. Test scores should be anchored to a concrete level of performance, such as lifting a certain number of pounds or running a specific distance within a certain amount of time. Minimum scores should be set consistent with the Secretary's commitment to not "reducing the qualifications for the job."

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

5. Implement Screening

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

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

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

6. Confirm Tests Are Working as Intended

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

Final Thoughts

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

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

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

Acknowledgments

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

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

Abbreviations

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

Chapter One. Introduction

On January 24, 2013, the Secretary of Defense and Chairman of the Joint Chiefs of Staff announced rescission of the 1994 Direct Ground Combat Definition and Assignment Rule (SecDef, 1994) and the intention to "integrate women into occupational fields to the maximum extent possible" (U.S. Department of Defense [DoD], 2013). The rule restricted assignments of women to occupational specialties or positions in or collocated with direct ground combat units below the brigade level, in long-range reconnaissance and special operations forces, and in positions including physically demanding tasks the "vast majority" of women cannot do (SecDef, 1994).

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

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

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

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

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

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

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

Why Standards?

What do we mean when we use the term *standards*? Throughout this report, we use *standards* or *physical standards* to refer to occupation-specific criteria that applicants must meet to enter or remain in a particular career field or specialty.[2] Some standards are valid for screening people for a particular job, and some are not. Standards that are *valid* are those that distinguish between those who are likely to be able to perform to the requirements of the job from those who are not. Standards that do a better job at making that distinction are more valid than those that do

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

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

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

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

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

---

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

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

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

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

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

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

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

> The chairman and the Joint Chiefs of Staff and I believe that we must open up service opportunities for women as fully as possible. And therefore today, General Dempsey and I are pleased to announce that we are eliminating the direct ground combat exclusion rule for women and we are moving forward with a plan to eliminate all unnecessary gender-based barriers to service.
>
> Our purpose is to ensure that the mission is carried out by the best qualified and the most capable servicemembers, regardless of gender and regardless of creed and beliefs. If members of our military can meet the qualifications for a job - and let me be clear, I'm not talking about reducing the qualifications for the job - if they can meet the qualifications for the job, then they should have the right to serve, regardless of creed or color or gender or sexual orientation.

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

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

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

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

Occupation-Specific Physical Standards and Entry into Military Service

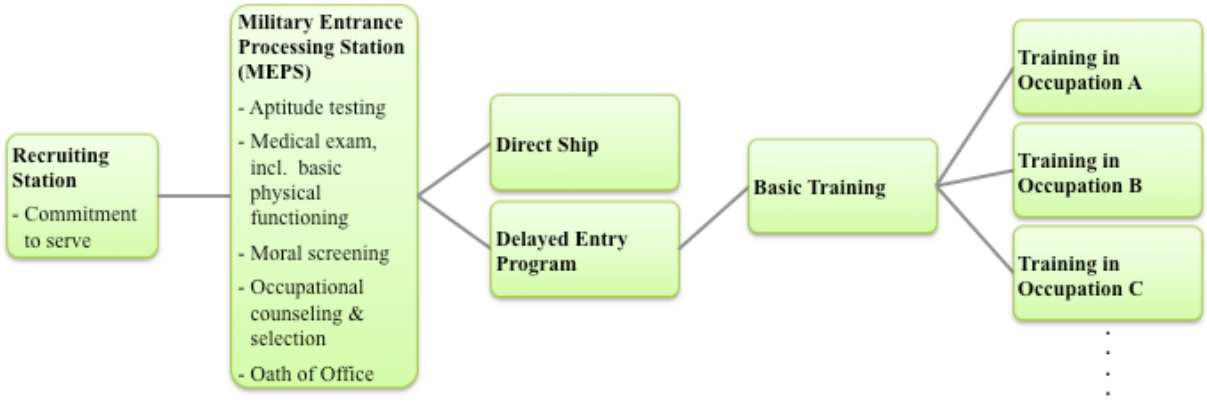
To develop appropriate tests and cutoff scores for screening and selecting new entrants for physically demanding military occupations, it is important to determine at which point in the entry process the screening will be done and standards for entry into the occupations determined. Figure 1.1 depicts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

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

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

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

Figure 1.1. Enlistment and Initial Training of Military Enlisted Personnel



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

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

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

Study Approach

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

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

Organization of This Report

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

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

Chapter Two. Methodological Approaches to Establishing Physical Job Requirements

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

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

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

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

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

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

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

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

Six-Stage Process

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

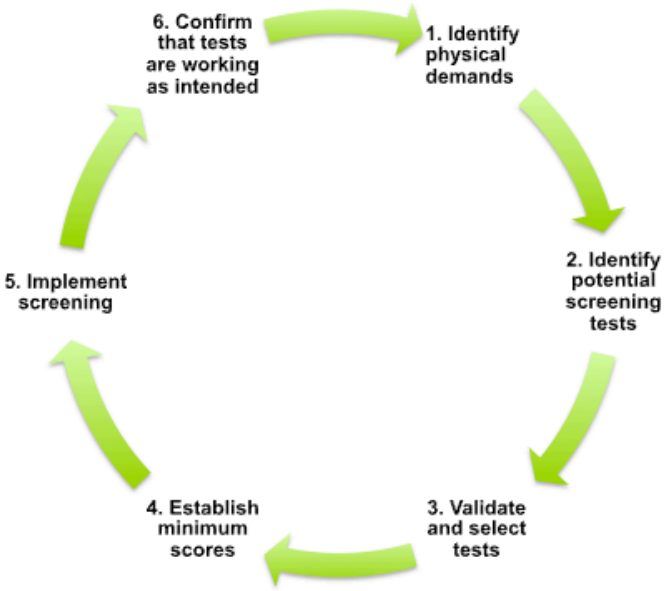
- **Stage 1. Identify the physical demands of the job.** Define all tasks required on the job, and identify which of those tasks are physically demanding and which are not. Identify other relevant aspects of performance, such as injuries, that may be affected by physical ability.
- **Stage 2. Identify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

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

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

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

Figure 2.1. Six Stages in Developing Physical Standards



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

Chapter Three. Identify the Physical Demands of the Job

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

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

Methods for Conducting a Job Analysis

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Reasons to Conduct a Careful Job Analysis

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

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

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

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

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

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

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

Chapter Four. Identify Potential Screening Tests

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

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

Taxonomies of Physical Aptitudes

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

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

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

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

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

Table 4.1. Fleishman's Physical Ability Domains

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

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

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

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

Examples of Tests Studied for Use in Employment Settings

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

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

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

Selecting Candidate Tests

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

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

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

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

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

---

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

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

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

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

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

Table 4.3. Example Considerations in Identifying Potential Screening Tests

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

Chapter Five. Validate and Select Tests

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

In the personnel selection context, the term *validate* has a much more precise meaning. It refers to the act of accumulating multiple sources of research-based evidence to support a test's use for a particular purpose (Messick, 1980, 1989, 1995; Anastasi, 1986; Binning and Barrett, 1989; 29 CFR Part 1607; 41 CFR Part 60-3; Society for Industrial and Organizational Psychology, 2003).[20] The following are three key types of validation evidence:

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

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

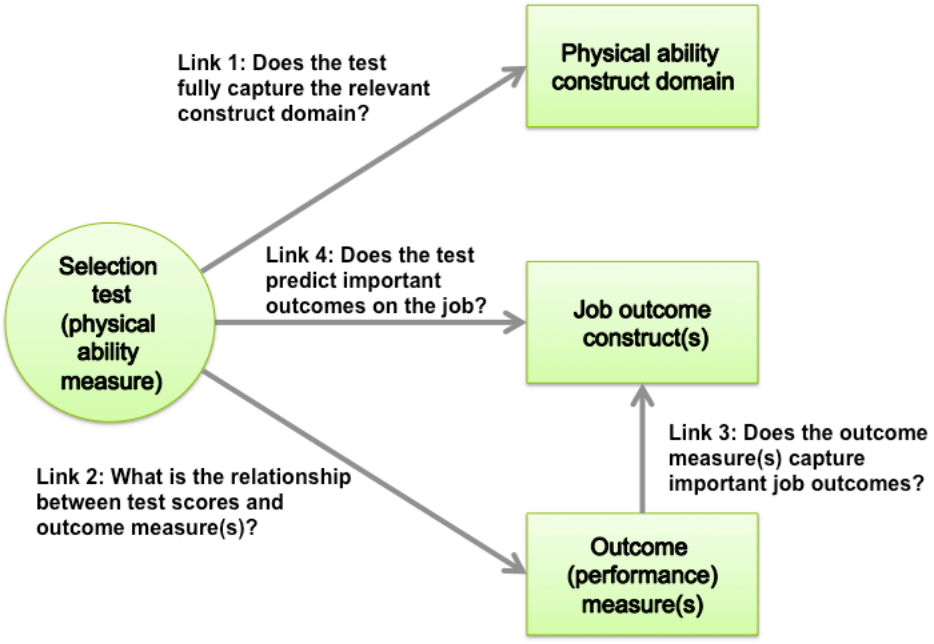
Figure 5.1 illustrates the conceptual linkages that can be examined during the validation process. In personnel selection, the ultimate goal of validation is to provide evidence to support link 4, that the selection test predicts important outcomes on the job. No single method of validation can provide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

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

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

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

Figure 5.1. Conceptual Validation Linkages



SOURCE: Adapted from Binning and Barrett, 1989.

Construct Deficiency and Construct Irrelevance

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

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

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

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

Figure 5.2. Construct Deficiency

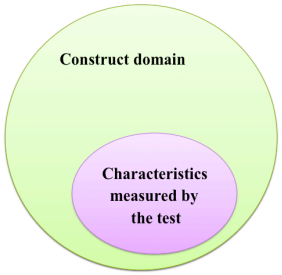
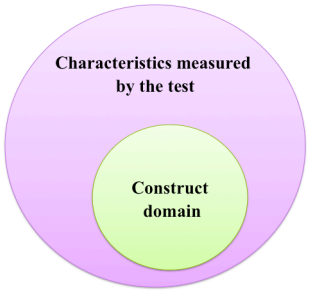


Figure 5.3. Construct Irrelevance



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

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

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

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

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

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

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

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

Content Validity

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

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

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

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

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

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

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

---

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

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

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

Criterion-Related Validity

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

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

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

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

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

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

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

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

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

and content analysis of the simulation support the simulation's overlap with key elements of the job, the findings would qualify as reasonable criterion-related validation evidence.

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

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

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

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

Convergent and Discriminant Validity

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

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

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

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

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

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

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

Fairness: Adverse Impact and Predictive Bias

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

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

---

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

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

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

Lay users of the terms often conflate *adverse impact* and *bias*, but, in personnel selection, these terms are not synonymous. In the personnel selection context, we are most concerned with a form of statistical bias known as *predictive bias*.[30] Predictive bias can take two forms. First, it can occur when predictive validity differs by group, a phenomenon known as *differential validity*. If the test is a better predictor of performance for one group than it is for another, then the test is considered biased against the group with the lower predictive validity.

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

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

---

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

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

Additional Considerations in Collecting Validation Evidence

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

Document the Process

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

Apply Appropriate Statistical Methods

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

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

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

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

Table 5.1. Example Considerations in Validating and Selecting Tests

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

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

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

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

Collect Multiple Sources of Evidence

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

---

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

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

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

Chapter Six. Establish Minimum Scores

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

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

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

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

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

Use of Job Analysis Data to Set the Minimum Score

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

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

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

Use of Expert Panels to Set the Minimum Score

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

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

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

Capture Expert Judgments About Minimum Performance on the Job

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

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

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

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

Capture Expert Judgments About Minimum Performance on the Test

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

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

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

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

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

Methods for Obtaining Expert Judgments for Setting Standards

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

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

A Well-Designed Study

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

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

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

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

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

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

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

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

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

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

Table 6.1. Example Considerations in Establishing Minimum Test Scores

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

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

Chapter Seven. Implement Screening

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

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

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

When the Test Should Be Administered

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

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

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

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

Standardize the Test Administration Procedures

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

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

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

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

---

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

Informing Applicants About the Test

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

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

Consider Phasing the Test in Gradually

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

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

Table 7.1. Example Considerations in Implementing Screening

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

Chapter Eight. Confirm that the Tests Are Working as Intended

Once initial standards for entry into physically demanding occupations are established, they will need to be the subject of ongoing research to regularly confirm that tests are working as intended. Even the best research designs leave some questions unanswered. New, unanticipated questions may arise after implementation. Some studies are feasible only after a test has been implemented. Constantly changing technology and mission can significantly alter the requirements of the job. For all these reasons, the research effort should be treated as an ongoing process, one that should continue long after a test has been implemented.

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

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

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

Institutionalize Research to Support Policy Changes

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

Reexamine Job Analyses

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

Continuously Collect Longitudinal Predictive-Validation Evidence

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

Reexamine Test Score Minimums

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

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

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

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

Evaluate Whether Administration Procedures Are Being Followed

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

Conduct Additional Research as Needed

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

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

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

Ongoing Personnel Research Efforts Are Not New

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

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

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

Chapter Nine. Final Thoughts

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

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

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

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

Once the standards have been developed, the last two stages of the six-stage process focus on implementation and sustainment:

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

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

We have provided an overview of the methods and data required to conduct each of the six stages for establishing standards for physically demanding jobs and identified key considerations at each stage. However, we have not addressed the many technical details involved. These details are determined based on the specific characteristics of the occupation, the environment, and any unique statistical needs or other issues encountered by the analysts. Carrying out the work requires expertise in a variety of domains, including industrial and organizational psychology, exercise physiology or a related field, psychometrics, and statistics. These experts rely on the expertise of SMEs from the occupation, who must be carefully selected to cover all types of work and work environments, and on appropriate test subjects drawn from the population of applicants, trainees, and job incumbents.

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

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

Phase II of the RAND Study

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

Glossary of Terms

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

assessment. See *test*.

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

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

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

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

construct contamination. See *construct irrelevant variance*.

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

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

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

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

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

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

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

cut score. See *standard*.

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

biased against the group with the lower predictive validity. Or it can occur when the predictive validity is equivalent for both groups but the test still underpredicts one group's performance relative to another group.

predictive validity. *Criterion-related validity* evidence that is collected as longitudinal data, i.e., data collected at two different times. Predictor information (data on the selection tests) is collected on personnel at time of hiring and outcome measures are collected after personnel have been on the job for some period of time. See *criterion-related validity*.

requirement. See *standard*.

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

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

selection test. See *test*.

simulation study. A validation study in which participants are measured on a predictor test, trained on how to perform key job activities, and tested on a series of simulations of those activities. If a relationship is shown between the test and the simulated outcomes and if job analysis data and content analysis of the simulation support the simulation's overlap with key elements of the job, the findings would qualify as reasonable criterion-related validation evidence.

standard. Used interchangeably with the terms *cut score* and *requirement*, *standard* refers to a criterion that an applicant must meet to enter or remain in an occupation. A minimum score on a physical test used to determine who is qualified for a job is one example of an occupation-specific entry standard.

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

task analysis. See *job analysis.*

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

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

tool. See *test*.

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

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

work analysis. See *job analysis*.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

DOL—*See* U.S. Department of Labor.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

SecDef—*See* Secretary of Defense.

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

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

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

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

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

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

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

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

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

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

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

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

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

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