PROJECT AIR FORCE

Evaluation of the Strength Aptitude Test and Other Fitness Tests to Qualify Air Force Recruits for Physically Demanding Specialties

The RAND Corporation

Sean Robson, Stephanie Pezard, Maria C. Lytell, Carra S. Sims, John E. Boon, Jr., Jason Michel Etchegaray, Michael Robbins, David Schulker, Jerry M. Sollinger, Jason H. Campbell, Anthony Atler, Stephan B. Seabrook

The Human Resources Research Organization

Deborah L. Gebhardt, Todd A. Baker, Erica K. Volpe, Kathryn A. Linnenkohl

Prepared for the United States Air Force

For more information on this publication, visit www.rand.org/t/RR1789

Library of Congress Cataloging-in-Publication Data is available for this publication.
ISBN: 978-0-8330-9942-6

Published by the RAND Corporation, Santa Monica, Calif.

Preface

The Air Force uses the Strength Aptitude Test (SAT) to determine whether recruits meet the fitness levels needed to perform the duties of various Air Force specialties with physical strength requirements. However, the SAT was developed in the early 1980s and has not been revalidated since then. In the interim, the duties associated with many Air Force Specialty Code (AFSC) classifications have changed, and new ones have been added. This report evaluates the status and validity of the SAT in a series of studies and builds upon previous RAND research on the SAT (Sims et al., 2014). It also suggests alternative strategies for developing SAT requirements that accurately reflect the physical demands of Air Force jobs while minimizing adverse effects on job opportunities for women.

The research reported here was commissioned by the Air Force's Force Management Policy Directorate (AF/A1P) and conducted within the Manpower, Personnel, and Training Program of RAND Project AIR FORCE. This report should interest Air Force leaders and staff concerned with standards to maintain the physical readiness of airmen who perform physically demanding tasks as part of their occupational specialty.

RAND Project AIR FORCE

RAND Project AIR FORCE (PAF), a division of the RAND Corporation, is the U.S. Air Force's federally funded research and development center for studies and analyses. PAF provides the Air Force with independent analyses of policy alternatives affecting the development, employment, combat readiness, and support of current and future air, space, and cyber forces. Research is conducted in four programs: Force Modernization and Employment; Manpower, Personnel, and Training; Resource Management; and Strategy and Doctrine. The research reported here was prepared under contract FA7014-16-D-1000.

Additional information about PAF is available on our website: www.rand.org/paf/

This report documents work originally shared with the U.S. Air Force on August 24, 2015. The draft report, issued on September 28, 2015, was reviewed by formal peer reviewers and U.S. Air Force subject-matter experts.

Contents

Figures

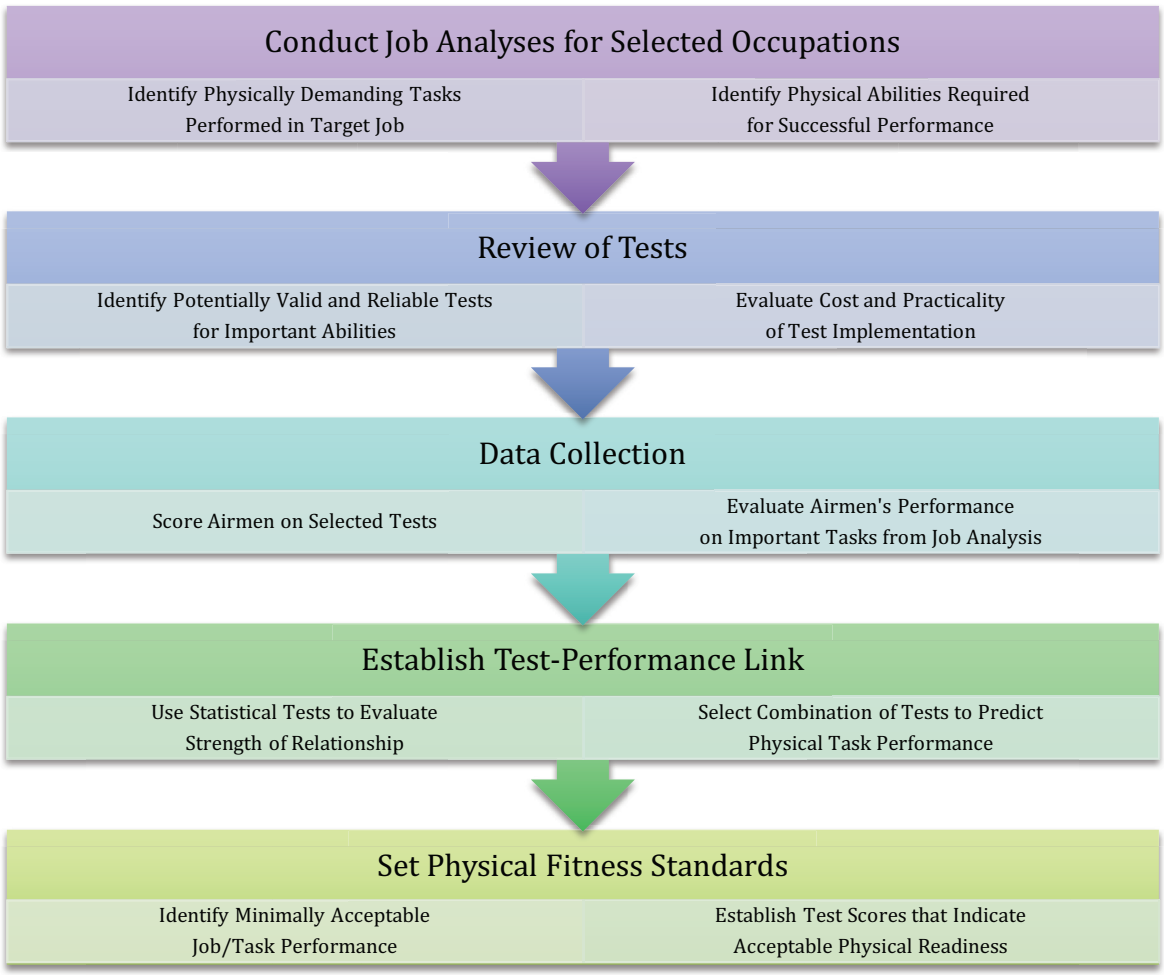
Tables

Summary

The Air Force wants to ensure that its recruits have the physical capability to perform the tasks of their duty positions, which can vary depending upon the specific demands of the position. To do so, the Air Force tests recruits' physical abilities as part of the induction process at the Military Entrance Processing Station (MEPS). Since the early 1980s, the Air Force has used the Strength Aptitude Test (SAT) to make this determination. The SAT is a weight-lifting test performed on an incremental lifting machine similar to equipment found in fitness centers. The test requires recruits to lift increasingly heavier weights until they either fail to lift the weight or they meet the weight requirement for their specific specialty.

But the composition of the Air Force has changed over time, as have the duties associated with the various occupational specialties. These changes require a reevaluation of the SAT's utility and effectiveness for qualifying recruits into these specialties. The Air Force asked RAND Project AIR FORCE to first evaluate potential benefits of the SAT and then develop and validate physical performance tests and standards to ensure airmen can perform the physically demanding tasks associated with selected enlisted Air Force Specialty Codes (AFSCs). To achieve these objectives, RAND conducted a series of studies between 2010 and 2015. These studies provide an initial evaluation of the SAT followed by job analyses and multiple validation efforts to determine whether the SAT and related fitness tests effectively indicate recruits' capabilities to perform physically demanding tasks required by AFSCs. Collectively, these studies provide the Air Force with scientifically based courses of action for implementing changes to ensure airmen can meet job-related physical requirements. This report summarizes the studies RAND has completed independently and one study conducted in conjunction with Human Resources Research Organization (HumRRO), which provided the additional data necessary to develop some courses of action for the Air Force to follow. A general outline for establishing test standards is presented in Figure S.1.

Figure S.1. Process for Establishing Test Standards



How Do Managers View the SAT?

RAND administered a survey to Career Field Managers (CFMs) to understand how they viewed the value of the SAT as an entry test and whether it should be continued. CFMs establish training, education, and related standards for the career fields they manage. Therefore, understanding their perspective is an important step in evaluating the potential advantages and disadvantages of the SAT. CFMs provided feedback in several areas, including the types of physical abilities required by the specialties they manage; whether the SAT requirements should be raised, lowered, or held constant; and benefits and challenges if the SAT were discontinued.

The survey responses indicated that the majority of CFMs are satisfied with current SAT requirements for the AFSCs they manage. Furthermore, CFMs identified more drawbacks than benefits if the SAT were eliminated. Although the CFMs perceived the SAT to play an important role in qualifying recruits for the AFSCs they manage, we concluded that further research should

address the validity of the SAT and evaluate the extent to which the SAT effectively predicts an individual's capability to perform the physically demanding tasks required by assigned AFSCs.

Does the SAT Predict Performance or Injuries?

RAND explored Air Force data from Enlisted Performance Ratings (EPRs) and work-related injuries to ascertain whether the SAT predicted either performance or susceptibility to injury. In our initial analyses, we found that available measures are largely insufficient for conducting the statistical tests needed to evaluate the relationships between SAT scores, performance, and injuries. For example, ratings on both the SAT and EPRs tend to cluster at the high end of the scale, which makes it difficult to identify any potential relationship. Also, changes in how the enlisted population is organized over time complicate the analysis, because some specialties get merged with others. Furthermore, SAT requirements (the minimum required for a given specialty) for some specialties have changed over time, and an individual's physical fitness can also vary over time, as evidenced by changes in SAT scores observed between week-zero and week-eight. Changes in SAT scores between MEPS and Basic Military Training (BMT) week-zero were also observed. With respect to injuries, the data contain very few, given the size of the population. We used injury data collected by the Air Force Safety Center, which may not capture less serious types of injuries for a variety of reasons, including policy guidance requiring base safety officials to conduct an investigation for injuries reported to the Air Force Safety Center (Copley et al., 2010), which may act as a disincentive to reporting less serious injuries. Given the limitations of existing data to evaluate the ability of the SAT to predict important job-related outcomes, we recommended a more comprehensive approach for identifying job-related physical requirements and potential physical fitness tests that could be used at the MEPS to determine the physical readiness of recruits to perform physically demanding job tasks associated with their assigned AFSC.

How Can Tests Be Linked to Physical Performance?

Given recent policy changes that open all assignments to women and the fact that the validity of the SAT has not been rigorously assessed since it was first developed, RAND, in conjunction with HumRRO, developed a methodology to deal with limitations of previous studies to examine the SAT's validity. Validation involves accumulating relevant evidence to provide a sound scientific basis for how tests, standards, training requirements, and related personnel decisions are applied. Although several strategies and sources of evidence can be used to establish validity, we evaluated the predictive validity of the SAT by conducting a concurrent, criterion-related validation study. This type of study helps to determine whether higher scores on the SAT are associated with higher physical task performance. In addition to the SAT, we also evaluated other physical tests to determine whether they would have higher validity or could be combined with the SAT to improve decisions about the level of fitness recruits need to perform physically

demanding tasks of a given AFSC. The study was designed to answer the following four questions:

1. What are the physical requirements to perform in different AFSCs?
2. How can physical performance on job-relevant tasks be measured?
3. Which physical fitness tests, including the SAT, indicate a recruit's capability to meet job-relevant physical demands?
4. Do the fitness tests predict physical performance equally well for different subgroups (e.g., men and women)?

The approach to answering these four questions consisted of the following tasks, executed jointly by RAND and HumRRO, primarily by HumRRO, or primarily by RAND:

Task 1: Identify specific tasks of selected AFSCs to identify the physical requirements to perform in different AFSCs. This task was executed jointly by RAND and HumRRO.

Task 2: Develop task simulations that approximate the types of physically demanding tasks performed across AFSCs. These task simulations measure physical performance across four movement patterns required to perform physically demanding tasks across AFSCs: (a) lifting and carrying, (b) lifting and holding, (c) climbing, and (d) pushing and pulling. This task was executed by HumRRO.

Task 3: Evaluate the predictive validity of physical fitness tests (for both men and women) to identify which tests can be used to indicate a recruit's capability to meet job-relevant physical demands. This task was executed primarily by RAND.

To accomplish the first task, RAND and HumRRO first analyzed the SAT data to identify career fields for analysis. That analysis showed that 38 percent of AFSCs require an SAT score of 40 pounds, and 26 percent require 70 pounds. Almost all of the men entering the Air Force lift 60 pounds or more, and about 87 percent of the women also do so. Furthermore, analysis of the scores suggests that the SAT begins to make a sizable difference for women at about 70 pounds, with almost all of the men and about 70 percent of women meeting this requirement. Taking these data into consideration along with data suggesting physical training from BMT can increase physical strength, emphasis was placed on the physical demands of AFSCs requiring a 70-pound SAT score or higher.

RAND and HumRRO interviewed CFMs and subject-matter experts for AFSCs requiring that score, asking them to identify the ten most physically demanding tasks and the level of that demand. Through these interviews, the physical demands representative of these AFSCs were identified to form the foundation for developing physical performance measures. Fitness tests were then evaluated by HumRRO to determine which ones could predict physical performance. Specifically, HumRRO identified nine fitness tests, including the SAT, for further analysis, and RAND conducted a series of analyses to develop several possible combinations of these tests (i.e., options) to strengthen the prediction of physical task performance. The tests, in addition to the SAT, are Arm Endurance, Arm Lift, Handgrip, Plank Test, Push-Ups, Sit-Ups, Standing Broad Jump, and Step Test. These are described in more detail in the main body of the report.

Each option had advantages and disadvantages. Some would require the purchase of relatively expensive equipment, some would have a greater adverse effect on job opportunities for women, and some offer no gains in validity. RAND assessed the following five options:

- Option 1: SAT is the only test used (baseline)
- Option 2: SAT plus any single test
- Option 3: SAT plus as many other tests as needed
- Option 4: SAT plus any single inexpensive test
- Option 5: SAT plus all inexpensive tests.

The results of the analysis indicate that adding the Arm Endurance test to the SAT adds the most validity of any test. The Arm Endurance test measures the ability of the muscles of the upper body to exert force repeatedly or continuously over a moderate time period. Thus, this test measures anaerobic power and muscular endurance. The test is conducted with a stationary arm ergometer, which resembles bicycle pedals but has handgrips instead of pedals. The individual "pedals" the ergometer with his or her hands for a minute and is scored on the number of revolutions achieved. Using it would require the purchase of an additional piece of equipment but would not require much additional space in the MEPS. It also reduces some of the potential problems of test bias, and it provides a sufficient increase in predictive validity to justify the additional costs of equipment and administering and scoring the tests.

Limitations

Although analyses consistently found support for the predictive validity of the SAT and the related fitness tests evaluated in the study, some significant limitations should be further addressed during a verification period before full implementation of any new tests or standards. Specifically, HumRRO explored options for recommending updated SAT standards for each AFSC; however, these efforts were unsuccessful due to limitations with the available data collected as part of the study. More specifically, HumRRO was unable to identify an acceptable algorithm to cluster AFSCs into meaningful groups (e.g., low vs. high physical demand) using the survey data collected by RAND. Alternative strategies to establish SAT cut scores were considered but could not be executed due to additional data that would be required from the Air Force specifying minimally acceptable job performance in each AFSC. Such data would allow the Air Force to establish a direct linkage between SAT standards and effective job performance; however, this type of data has not yet been collected by the Air Force. In consideration of these data limitations, RAND provides several courses of action, all of which require maintaining the current standards until additional data can be collected to establish the SAT scores associated with minimally acceptable performance within each AFSC.

Courses of Action (COAs) the Air Force Could Pursue

The research done for this study indicates that the SAT remains a valid measure of a recruit's ability to perform the physical duties of his or her Air Force specialty. However, augmenting the SAT with additional physical test(s) could increase the validity of the testing done at the MEPS. Alternatively, the Air Force could continue administering only the SAT at the MEPS and shift the final determination of physical capabilities to perform the duties of a given AFSC to training (rather than entrance) standards. For each of the COAs, RAND considered several factors, including resource requirements (e.g., costs), how well fitness test scores correlate with performance (i.e., validity), and potential gender test bias. Gender test bias can occur in several ways and, depending on the nature of the bias, test scores may not be a good indicator of a particular subgroup's performance. In the context of physical fitness testing, the presence of test bias could mean a greater proportion of one subgroup (e.g., women) is classified into a specialty for which members cannot perform the physical tasks to an acceptable level. The four COAs we analyzed are as follows:

COA #1. Adopt the physical test battery at the MEPS that maximizes validity. The combination of tests that meets this objective includes the SAT, Arm Endurance, Push-Ups, and Handgrip.

COA #2. Adopt a physical test battery at the MEPS that maximizes validity with no additional equipment costs; combines Standing Broad Jump with SAT.

COA #3. Adopt a physical test battery at the MEPS that maximizes validity with limited additional costs; combines SAT with Arm Endurance test.

COA #4. Retain the SAT as the only physical test at the MEPS.

The analysis of the four courses of action appears in Table S.1.

Table S.1. Advantages and Disadvantages for Each COA

COA	Advantages	Disadvantages
COA #1. Adopt the physical test battery at the MEPS that maximizes validity. The combination of tests that meets this objective includes the SAT, Arm Endurance, Push-Ups, and Handgrip.	• Maximizes potential to ensure recruit has the ability to perform physically demanding tasks • Provides the most comprehensive assessment of physical fitness, to include combinations of tests measuring muscular strength and muscular endurance • No gender test bias indicated	• Requires additional resources and costs for Handgrip and Arm Endurance • May have time and space implications for MEPS • Return on investment diminishes for each additional test • Evidence on how to combine test scores is limited
COA #2. Adopt a physical test battery at the MEPS that maximizes validity with no additional equipment costs. Combines Standing Broad Jump with SAT.	• Increases validity beyond the SAT with a test that requires no additional costs and minimal resources to administer	• Gains in validity over the SAT (+4%) minimal and likely do not justify cost and additional resources to administer • Adding in all other no-cost tests still offers limited validity gains over the SAT (+7%) • Test may overpredict female performance and underpredict male performance on tasks (potential gender test bias)
COA #3. Adopt a physical test battery at the MEPS that maximizes validity with limited additional costs. Combines SAT with Arm Endurance test.	• Balances cost and validity gains • Validity increases significantly beyond the SAT (+22%) • Involves fewer tests • Reduces gender test bias compared with using SAT alone	• Slightly less validity gain than COA #1 • Increases costs somewhat for equipment, maintenance
COA #4. Retain the SAT as the only physical test at the MEPS.	• Requires only the SAT test and takes advantage of the relatively strong correlation with physical task performance • Requires minimal changes at MEPS	• Slightly less validity gain than other COAs • Potential gender test bias

Implementing a COA

Given the study limitations and potential effect on each AFSC, RAND recommends maintaining the SAT requirements currently in place while following an implementation plan to verify any COA selected by the Air Force. Specifically, we recommend the following steps:

1. Integrate job analysis physical demand survey items into Occupational Analysis Division's routine surveys of each AFSC. The survey items discussed in Chapter Four can be used. Responses to survey items should be evaluated for differences across subgroups (e.g., location, gender). Periodically verify the accuracy of responses (e.g., weight of equipment) by referencing official documents on the dimensions and weights of equipment, and by directly observing and weighing equipment during site visits.
2. Provide CFMs and other senior leaders in each AFSC with the SAT requirements summary job analysis data for the AFSCs they manage.

3. Collect feedback and address questions or concerns from CFMs and other senior leaders regarding job analysis survey results.
4. Begin administering any new test(s) (e.g., Arm Endurance) at the MEPS to gather data on new Air Force recruits.
5. Collect data on physical performance of recruits assigned to each AFSC.
6. Use the test data collected from the MEPS and the physical performance data to verify the accuracy of the SAT requirement and to identify other test scores (i.e., requirements) associated with minimally effective task performance for each AFSC.
7. Calibrate and adjust requirements based on feedback and data collected.
8. Establish a system for regular monitoring and updating of test requirements.

RAND recommends that CFMs, Training Pipeline Managers, and Training Cadre review the results from the job analysis survey to identify critical physical tasks that can serve as a foundation for physical standards in technical training (i.e., used in physical task simulations). RAND also recommends implementing a feedback system to monitor whether trainees are meeting these standards. If a certain percentage of trainees (e.g., greater than 5 percent) cannot meet standards, that should trigger a review of the SAT standards for that AFSC. If the SAT requirement is found to be acceptable, an additional physical demands study conducted by the Air Force Fitness Testing and Standards Unit should be initiated. This study should examine the physical requirements of the AFSC and consider whether additional physical ability screening is required during the recruitment phase.

The Air Force may wish to consider whether concentration of physical testing resources to the most demanding occupations would enable their most efficient deployment regardless of the COA chosen. As described in this report, only a subset of AFSCs have physical requirements; therefore, focusing efforts on those AFSCs with the greatest physical demands should result in more fidelity and greater efficiency in the overall process. Finally, the COAs described all include development of a system to ensure that the Air Force continues to update physical requirements along with changes in the Air Force jobs themselves, which is key to maintaining the validity of those requirements and, hence, key to ensuring the requirements are beneficial.

Acknowledgments

The study team would like to thank its sponsors, which have included General Gina Grosso and General Brian Kelly (AF/A1P). We would also like to thank Dr. Lisa Hughes (AF/A1PT) for providing direction, feedback, and support throughout the study. Lt Col Charles Parada (AF/A1) was instrumental in coordinating the validation study, particularly with his diligent leadership in working with BMT and Technical Training to identify potential research participants, identify medical monitors, and ensure the project could be completed on time. John Trent (AFPC/DSYX) was also critical to the overall success of the validation study. He took on much of the responsibility to secure a location for the study, identify and procure equipment, and coordinate the design and building of the task simulations. Without the hard work and dedication of both Lt Col Parada and John Trent, this study would not have been possible. This study was also supported by David Crane, who, with his team, constructed platforms and related equipment for the validation study. We also thank Hector Acosta (AF/AFRS) and Dr. Bruce Burnham (HQ Air Force Safety Center) for providing archival data to evaluate potential links between SAT scores and injuries. Brian Chasse (AF/AFPC) provided data on the SAT retesting.

We would also like to thank our RAND colleagues who provided the necessary support for administering tests during the validation study, including a combination of graduate students from the Pardee RAND Graduate School (Gursel Aliyev, Aziza Arifkhanova, Carlos Buiterrez, Therese Jones, Beth Katz, Katie Loa, Nelly Mejia, Claire O'Hanlon, and Steve Trochlil) and research assistants (Amy Grace Donohue, Molly Doyle, Daniela Kusuke, Chris Maerzluft, and Laura Raaen). Paul Emslie provided data to support the identification of technical training locations. Tom Brogden and Karin Lui from RAND's Survey Research Group helped design and execute the online administration of the physical demands survey. Interviews with the career field managers (CFMs) were conducted by Anthony Atler, Jason H. Campbell, Abby Robyn, and Stephen B. Seabrook. James Gazis designed a website to communicate the purpose of the research study and allow interested Air Force personnel to sign up to receive additional information on how to participate in the study.

Finally, we offer special thanks to all of the Career Field Managers for their time and support for the interviews, site visits, and equipment provided for the study; and a great deal of gratitude to all the Air Force personnel who participated in one or more phases of the validation study.

Abbreviations

AIC	Akaike's information criterion
AFB	air force base
AFECD	Air Force Enlisted Classification Directory
AFPC	Air Force Personnel Center
AFS	Air Force Specialty
AFSC	Air Force Specialty Code
BMT	Basic Military Training
COA	course of action
CFM	Career Field Manager
DCAR	Direct Ground Combat Definition and Assignment Rule
EOD	Explosive Ordnance Disposal
EPR	Enlisted Performance Ratings
HumRRO	Human Resources Research Organization
HPS	Human Performance Systems
ILM	incremental lifting machine
MCQ	Movement Classification Questionnaire
MEPS	Military Entrance Processing Station
OAR	Occupational Analysis Report
PPE	personal protective equipment
POC	point of contact
SAT	Strength Aptitude Test
SERE	Survival, Evasion, Resistance and Escape
SD	standard deviation
SME	subject-matter expert

Chapter One. Introduction

Background on the SAT

Since the 1980s, the Air Force has used a physical strength test known as the Strength Aptitude Test, or SAT, to ensure recruits have the physical strength to perform the physical demands of their Air Force specialties. The SAT is a weight-lifting test performed on an incremental lifting machine (ILM) similar to the equipment found in fitness centers. The Air Force administers the SAT to recruits at Military Entrance Processing Stations (MEPSs). The SAT involves lifting a series of weights on the ILM from six inches above the ground to a height of six feet. The initial weight lifted is 40 pounds. If an individual successfully lifts this weight, 10 pounds are added to the ILM, and a lift is performed with 50 pounds. The weight is increased 10 pounds at a time until the individual is no longer able to lift to six feet or attains a score of 110 pounds, the current allowable limit. The score for the test is the heaviest weight lifted successfully to six feet. The SAT and other ILM tests have been shown to be related to job tasks involving manual materials handling (Ayoub et al., 1987; Gebhardt, Baker, and Thune, 2006; Knapik et al., 2004; Myers, Gebhardt, and Crump, 1984; Teves, Wright, and Vogel, 1985).

Job qualification standards on the SAT have been established for all Air Force specialties. Originally, the standards were developed by computing an average physical demand weighted by frequency of performance and percentage of the Air Force Specialty (AFS) members performing a task (McDaniel, Skandis, and Madole, 1983).

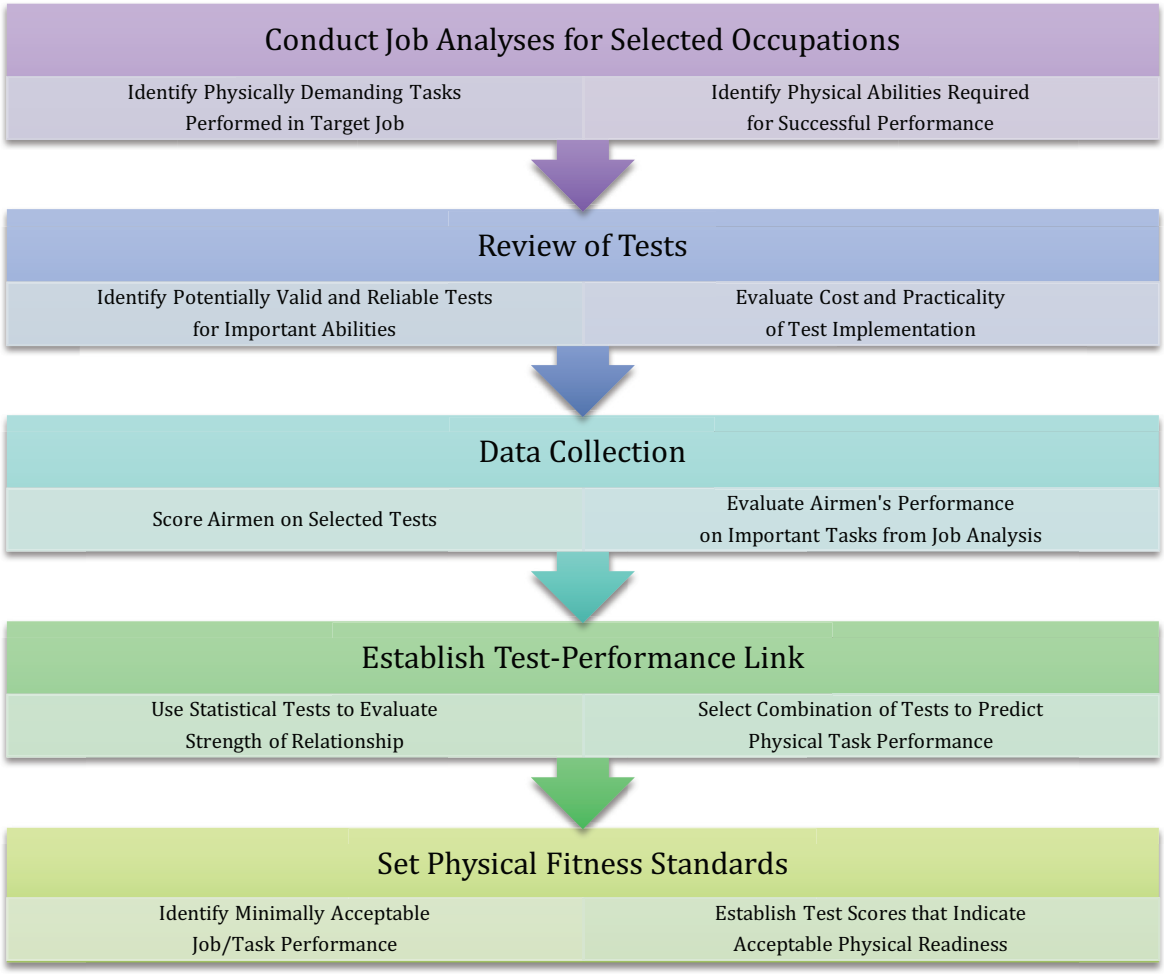
Air Force specialties were originally surveyed for development of SAT standards between 1978 and 1982 (McDaniel, Skandis, and Madole, 1983). During the mid-1990s, the U.S. General Accounting Office (GAO) reviewed the use and development of gender-neutral occupational performance standards in the Department of Defense (DoD) (GAO, 1996). Among other things, the GAO questioned the effectiveness of the SAT in predicting capability to do physically demanding tasks, noting problems in the administration of the test to new recruits and delays in updating occupational requirements.

RAND's SAT Studies

Given the issues outlined by the GAO and specific objectives to evaluate the utility and validity of the SAT, the Air Force asked RAND to develop and validate physical performance tests and standards—including those for the SAT—that should be used to ensure airmen are capable of performing physically demanding tasks associated with selected enlisted Air Force Specialty Codes (AFSCs). To achieve these objectives, RAND conducted a series of studies from fiscal years 2010 through 2016. These studies provide an initial evaluation of the SAT, followed by job analyses and multiple validation efforts to determine whether the SAT and

related fitness tests are effective indicators of recruits' physical capabilities to perform physically demanding tasks required by Air Force occupational specialties. These studies collectively provide the Air Force with scientifically based courses of action for implementing changes to ensure airmen are capable of meeting job-related physical requirements. A general outline for establishing test standards is presented in Figure 1.1.

Figure 1.1. Process for Establishing Test Standards



The first of the RAND SAT studies (conducted from fiscal years 2009 to 2012) (Sims et al., 2014) concluded that there were some inconsistencies in test administration at the MEPSs and that the process for setting strength requirements for AFSCs might be deficient because it involves only limited input from information gathered from site visits to three locations for each AFSC. The report offered recommendations for addressing inconsistencies in test administration at the MEPSs and improving inputs in setting strength requirements. In addition, the study recommended that the Air Force use an alternate method to convert job demands into SAT

requirements and that it collect data on the SAT and other physical tests before and after basic training to support future validation efforts.

The studies described in this report address recommendations outlined in a 2014 RAND report (Sims et al., 2014), as well as concerns outlined by the GAO (1996) report. The studies described here were conducted between fiscal years 2013 and 2016 and have two overarching purposes. First, they evaluate the status and predictive validity of the SAT. RAND first attempted to address this question by linking SAT scores to existing data in Air Force databases. These efforts revealed that the existing data were insufficient to draw conclusions about the validity of the SAT. Therefore, RAND partnered with Human Resources Research Organization[1] (HumRRO) to design and conduct a more comprehensive study of the SAT by collecting and evaluating new data. HumRRO's primary roles were to conduct job analyses to identify job-specific physical requirements, develop realistic task simulations to approximate the physical demands of Air Force occupational specialties, identify potential tests for measuring the physical abilities needed to perform physically demanding job tasks, and design and collect data using a criterion-related validation study design.

The data collected from the criterion-related validation study provided the foundation for addressing the second overarching objective of this report: to suggest alternative strategies for developing SAT requirements (that is, scores below which a recruit does not qualify to enter a given AFSC) to reflect accurately the physical demands of Air Force jobs while minimizing adverse effects on job opportunities for women. To address this objective, RAND used data collected by HumRRO to explore the links between the SAT, related physical fitness tests, and performance on job-related physical tasks (i.e., physical task simulations). Building on the findings from these analyses, RAND also evaluated different options for combining fitness tests to meet different organizational objectives (e.g., high validity–low cost).

Methodology of Studies in This Report

RAND and HumRRO used several strategies to develop scientifically based courses of action:

1. reviewed and evaluated existing validation evidence (RAND)
2. surveyed Career Field Managers (CFMs) to identify their position on the adequacy and utility of the SAT standards (RAND)
3. evaluated the pass rates at different SAT cut points (RAND)
4. conducted job analyses to identify job-specific physical demands (RAND and HumRRO)
5. evaluated criterion-related validity of SAT on the following:
    a. job performance (RAND)

[1] HumRRO acquired Human Performance Systems during the final stages of the study; therefore, we reference a partnership with HumRRO even though the original partnership was established with Human Performance Systems.

b. injury risk (RAND)

6. designed and conducted a criterion-related validation study to measure performance on the SAT, related physical fitness tests, and job-related physical tasks (HumRRO)
7. evaluated the predictive validity of the SAT and different tests and developed different options for combining fitness tests to meet different organizational objectives (e.g., high validity–low cost) (RAND).

Organization of This Report

This report first presents results from studies independently conducted by RAND (Chapters Two and Three). Chapter Two discusses the benefits and challenges of keeping (or removing) the SAT according to a sample of CFMs. Chapter Three presents RAND's initial efforts to establish the validity of the SAT using existing data available in Air Force databases. The following three chapters (Chapters Four through Six) present the planning, design, execution, and analysis of data from the criterion-related validation study jointly conducted by RAND and HumRRO. More specifically, Chapter Four discusses the objectives of the criterion-related validation study and its overarching methodology. Chapter Five primarily focuses on the specific details of HumRRO's design and data collection for the criterion-related validation study. Chapter Six presents RAND's analysis of the data HumRRO collected in the criterion-validation study to evaluate the linkages between the SAT, related physical fitness tests, and job performance.

A concluding chapter reviews several courses of action for adopting updated tests and standards and offers a series of steps to consider during an implementation period to conduct additional review and evaluate proposed changes. Following these steps will help the Air Force meet its ultimate goal to establish a system for identifying and regularly updating occupation-specific physical demands and the corresponding associated tests and standards for screening airmen into physically demanding occupations.

Chapter Two. Manager Views of Benefits and Challenges of SAT

This chapter describes the results of a survey (see Appendix A) RAND administered to CFMs to understand how they viewed the value of the SAT as an entry test and whether it should be continued. CFMs establish training, education, and related standards for the career fields they manage. Furthermore, SAT scores are reexamined at the request of CFMs. Following a standard protocol evaluating the physical tasks performed by the AFSC, an algorithm is used to generate an SAT estimate. The SAT estimate is then reviewed by the CFM, who may request an adjustment to the SAT requirement to better reflect the physical requirements of the AFSC. A more detailed description of this SAT standard-setting process in provided in Sims et al. (2014). Given the role of CFMs in setting standards for the SAT, understanding their perspective is an important step in evaluating the potential advantages and disadvantages of the SAT.

In addition to understanding the CFMs' perspective, the survey also attempted to identify the range of physical demands required by occupational specialties in the Air Force to better understand the relevance of the SAT and potential need for other physical screening tests. Although specific AFSCs require strength, several other physical abilities (e.g., aerobic endurance) may also be needed to perform job-specific duties. The chapter begins by describing the breadth of coverage for AFSCs the survey provided and whether CFMs thought the SAT requirements should be raised, lowered, or held constant. They were also polled on the requirements of the specific career fields they manage and were asked to identify benefits and challenges if the SAT were discontinued.

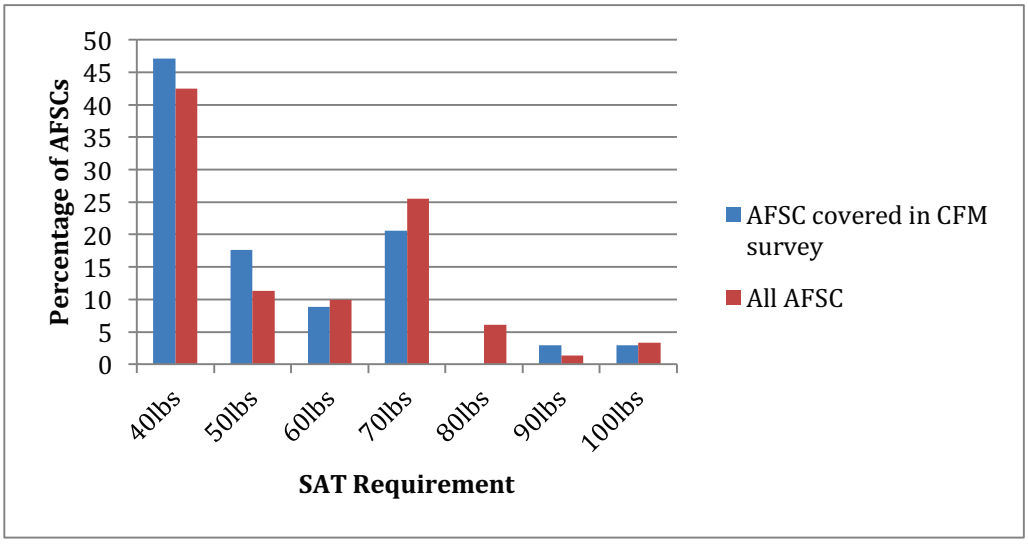
Survey Responses

Using a contact list provided by the Air Force, we invited all CFMs in the Air Force to complete the survey. Twenty-nine out of the 70 CFMs responded, yielding a 41-percent return rate. Some CFMs answered questions on more than one AFSC. Consequently, results cover a total of 34 AFSCs. The distribution of SAT requirements for AFSCs included in this set of responses followed closely the distribution for all AFSCs in general, except for the 80-pounds category, which is not represented in our survey results. Figure 2.1 shows by percentage the weight requirement for a recruit to qualify for an AFSC.[2] The most represented category of

[2] The numbers of AFSCs covered in our survey, compared with the total number of AFSCs, are the following for each SAT requirement: 40 pounds—16 AFSCs covered in our survey, out of a total of 90 AFSCs; 50 pounds—six AFSCs, out of a total of 24; 60 pounds—three AFSCs out of a total of 21; 70 pounds—seven AFSCs, out of a total of 54; 80 pounds—0 AFSCs, out of a total of 13; 90 pounds—one AFSC, out of a total of three; 100 pounds—one AFSC, out of a total of seven.

AFSCs, both in our sample and in the general AFSC population, also has the lowest requirement: 40 pounds.

Figure 2.1. Physical Requirements Distribution, by Percentage

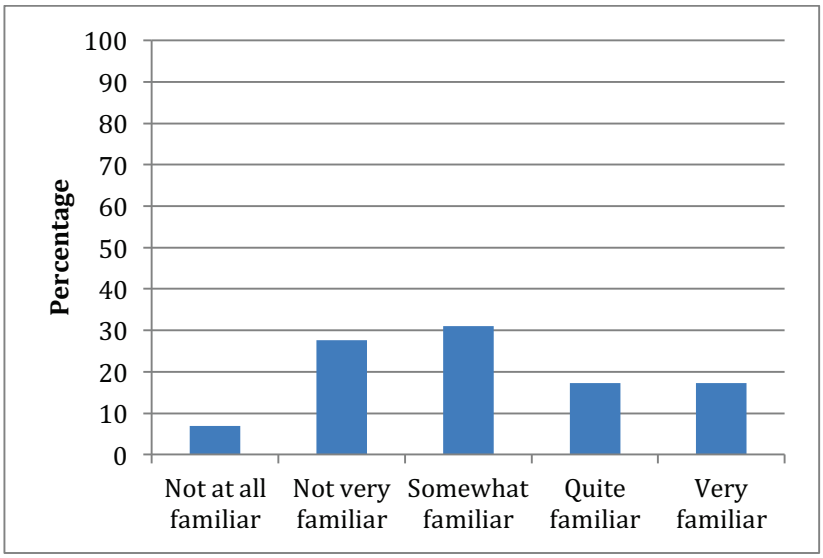


NOTE: SAT requirements are based on the *Air Force Enlisted Classification Directory*'s (AFECD's) Mandatory AFSC Entry Requirements Table (AFECD, 2013), which was current at the time of this study.

Figure 2.2 shows the CFM respondents' reported levels of familiarity with the SAT. Most respondents claim some familiarity with the SAT: Only two respondents said they were "not familiar at all" with it. However, familiarity was limited, because 65 percent of respondents were "not very familiar" or "somewhat familiar" with the SAT, while 34.5 percent were "quite familiar" or "very familiar" with it. Four respondents gave incorrect requirements for their AFSCs, in spite of having been provided with the Mandatory AFSC Entry Requirements Table that contains this information along with the survey. Of these four respondents, who were off by 4 to 20 pounds, one claimed to be "quite familiar" with the SAT (see Figure 2.2).[3]

[3] Three of these respondents thought the SAT requirement for their AFSC was higher than it actually is; one respondent thought it was lower.

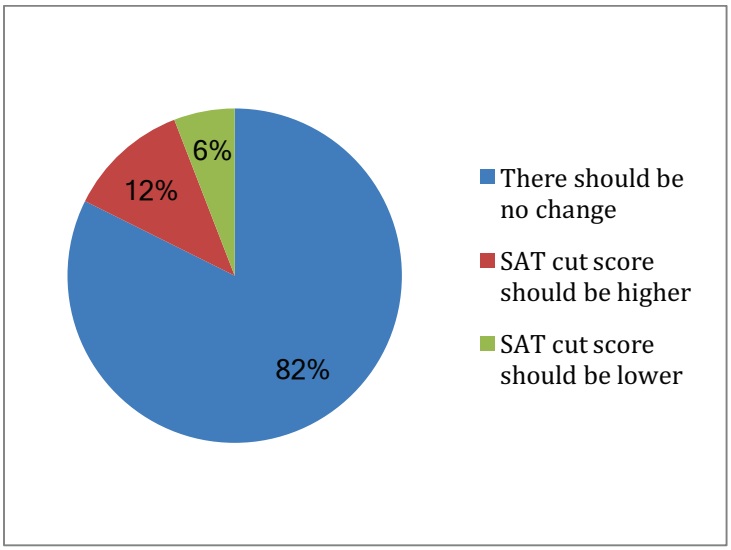
Figure 2.2. Level of Familiarity with SAT



NOTE: Numbers of respondents: "Not at all familiar," two respondents; "Not very familiar," eight respondents; "Somewhat familiar," nine respondents; "Quite familiar," five respondents; "Very familiar," five respondents.

Overall, most respondents favored keeping the SAT cut score as it is for their AFSCs, with 82 percent of the AFSCs covered in our sample deemed to have the right SAT cut score (see Figure 2.3). Those respondents who advocated a change proposed a minimal one (usually ten pounds; 20 pounds on two occasions). In four instances, respondents called for an increase in the SAT cut score because job equipment is now heavier or because the mission changed and is now more demanding. In two instances, one respondent asked that SAT cut score be lowered for two AFSCs to match the requirement of two other AFSCs with similar physical demands.

Figure 2.3. CFM Opinions Regarding Changes to SAT Requirements



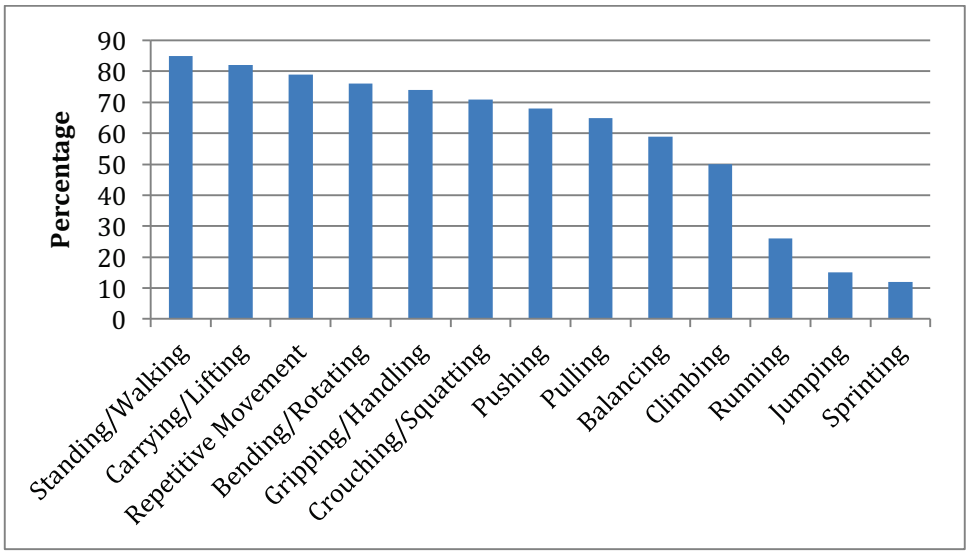
NOTE: Numbers of respondents: "There should be no change," 28 respondents; "SAT requirement should be higher," four respondents; "SAT requirement should be lower," two respondents.

Those respondents who had mentioned being "quite familiar" or "very familiar" with the SAT were more likely to advocate for the SAT remaining the same (90 percent, compared with 79 percent for respondents "not at all familiar," "not familiar," or "somewhat familiar" with the SAT). None of these respondents advocated lowering the SAT cut score, while 8 percent of those least familiar with the SAT offered that recommendation.[4]

CFMs were also asked a general question about the physical requirements of the AFSCs they manage. Figure 2.4 shows the physical movements that CFMs mentioned as being required in the AFSCs covered in our survey.

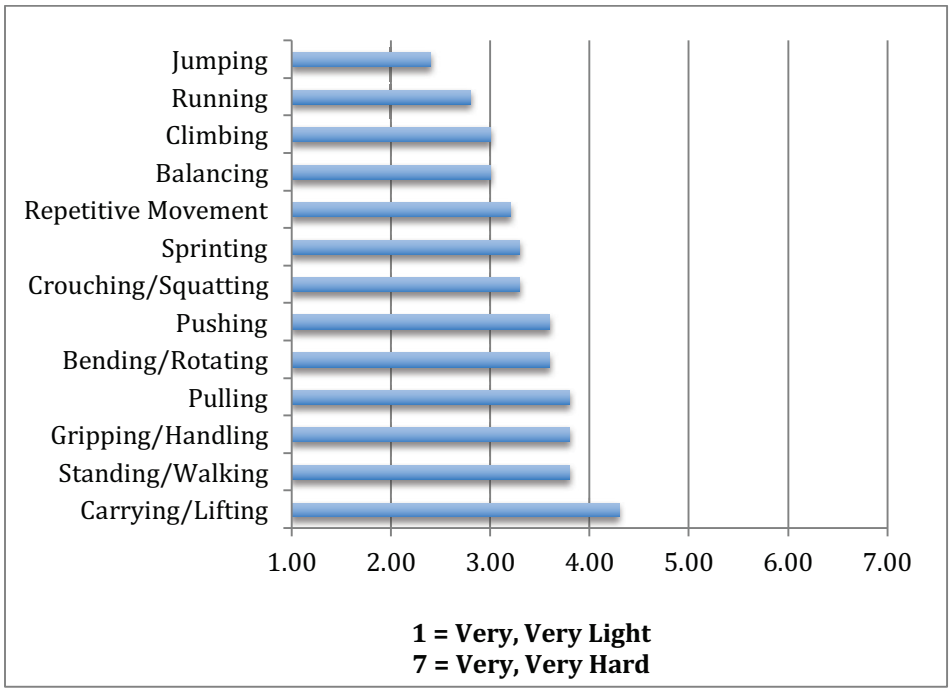
[4] N=10 for respondents "quite familiar" or "very familiar" with the SAT; N=24 for respondents "not at all familiar," "not familiar," or "somewhat familiar" with the SAT. Note that respondents (N=2) who are CFMs for several AFSCs are counted multiple times—once for each AFSC they oversee, since they may have given different answers as to whether the requirements for these different AFSCs should remain the same or not.

Figure 2.4. Physical Movements Required by AFSCs, by Percentage of AFSCs Covered in Survey



The physical movements most often required by the AFSCs covered in our survey are standing/walking (85 percent) and carrying/lifting (82 percent). These are also some of the most demanding physical movements in terms of level of effort as reported by CFMs. The average effort rating is highest for carrying/lifting (average of 4.30 on a one-to-seven scale, with 1 representing "Very, Very Light" and 7 representing "Very, Very Hard") followed by standing/walking, gripping/handling, and pulling (average of 3.80 each) (see Figure 2.5).

Figure 2.5. Average Levels of Effort for Physical Movements



NOTE: Averages were calculated based on the responses of 29 CFMs representing 34 AFSCs.

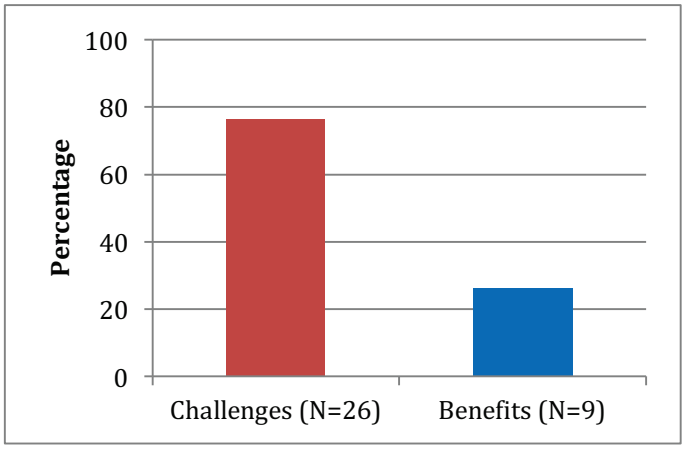
Benefits of and Challenges to Removing the SAT

A comparison between the number of respondents who cited at least one challenge to removing the SAT (that is, a problem that would occur if the SAT were eliminated) and those who cited at least one benefit to doing so suggests that perceptions of challenges loom larger than perceptions of benefits (positive outcomes if the SAT were eliminated), even when taking into account the fact that the survey offered more suggestions for challenges and negative outcomes (five categories: increased attrition, decreased job performance, increased risk of injuries, decrease in efficiency, and other challenges) than for benefits (three categories: increased manning, increased opportunity, and other benefits) (see Figure 2.6).

Out of 29 respondents, nine cited "other challenges." They cited risk to performance on four occasions, even though this category was mentioned in the multiple-choice part of the question, and three out of these four respondents had already selected it. They included risk to safety, the need for a physical standard, risk of damage to equipment, and the fact that removing the SAT may lead to longer training time. By comparison, only one respondent cited an "other benefit"—"Removing something useless"—of removing the SAT. While all those respondents who had mentioned being "quite familiar" or "very familiar" with the SAT cited at least one challenge,

only 67 percent of those respondents "not at all familiar," "not familiar," or "somewhat familiar" with the SAT did.[5]

Figure 2.6. Percentage of Responses Indicating Challenges and Benefits to Removing SAT

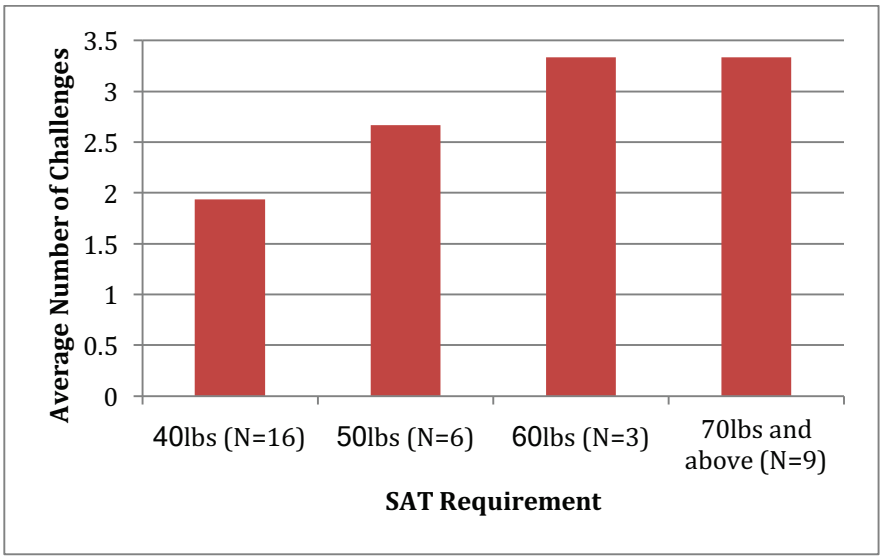


As shown in Figure 2.7, more potential challenges are identified by CFMs managing AFSCs with higher physical requirements, although the level of potential challenges seems to plateau after 60 pounds. No major differences appear among the CFMs in how they rank challenges except for those managing AFSCs with 40-pounds requirements, who tend to emphasize increased risk of injuries more, and CFMs managing AFSCs with requirements of 70 pounds and above, who tend to identify increased attrition as a challenge less frequently. No major differences appear between the CFMs who are most familiar with the SAT (2.7 challenges cited on average) and those least familiar (2.5 challenges cited on average).[6]

[5] N=10 for respondents "quite familiar" or "very familiar" with the SAT; N=24 for respondents "not at all familiar," "not familiar," or "somewhat familiar" with the SAT. Note that respondents (N=2) who are CFMs for several AFSCs are counted multiple times—once for each AFSC they oversee, since they may have given different answers as to whether there would be challenges or benefits to removing the SAT for these different AFSCs.

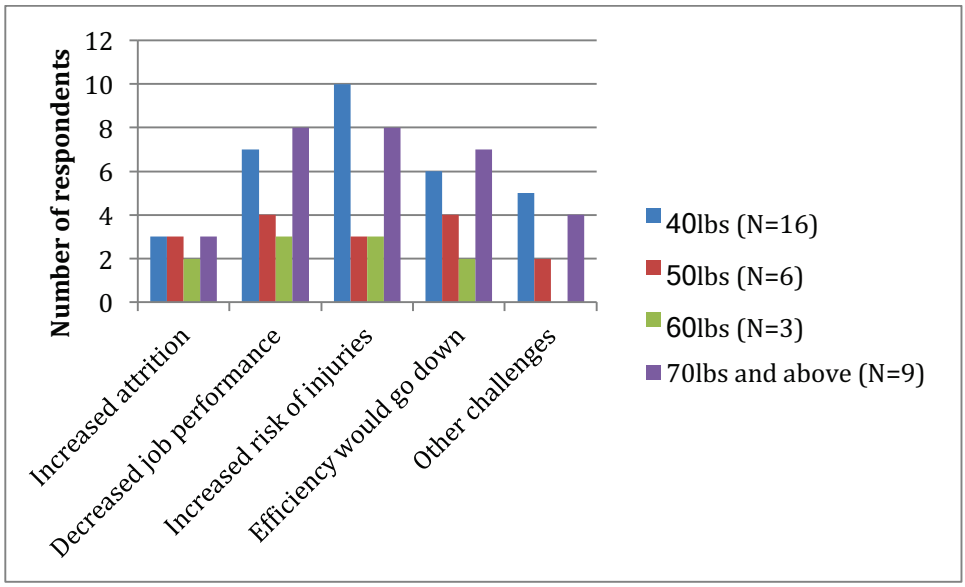
[6] N=10 for respondents "quite familiar" or "very familiar" with the SAT; N=24 for respondents "not at all familiar," "not familiar," or "somewhat familiar" with the SAT. Note that respondents (N=2) who are CFMs for several AFSCs are counted multiple times—once for each AFSC they oversee, since they may have given different number of challenges for these different AFSCs.

Figure 2.7. Average Number of Challenges Cited, by Requirement



The most-cited drawbacks to removing the SAT are increased risk of injuries, decreased job performance, and reduced efficiency. As shown in Figure 2.8, respondents from AFSCs with the lowest requirements were more likely to mention increased risk of injuries above other concerns. In the 70 pounds and above category, increased risk of injuries and decreased job performance are cited the same number of times as the paramount concern for removing the SAT.

Figure 2.8. Challenges to Removing SAT Cited, by Requirement



NOTE: Respondents (N=2) who are CFMs for several AFSCs are counted multiple times—once for each AFSC they oversee, since they may have cited different challenges for these different AFSCs.

Summary of CFM Survey

The survey of CFMs indicated that the majority is satisfied with current SAT requirements for the AFSCs they manage. Furthermore, CFMs identified more drawbacks than benefits if the SAT were eliminated. These results should be interpreted cautiously, since only 34.5 percent of the CFMs felt “quite” or “very familiar” with the SAT. Although the CFMs perceived the SAT to play an important role in qualifying recruits for the AFSCs they manage, we concluded that further research should address the validity of the SAT and evaluate the extent to which the SAT effectively predicts an individual’s capability to perform the physically demanding tasks required by the relevant AFSC. The following chapter describes our initial attempts to evaluate the validity of the SAT.

Chapter Three. The Validity of SAT Scores

An important piece of evidence about the validity of any employment test is whether that test can predict important job-related outcomes. This type of validity (known as criterion-related validity) is particularly important in physical ability testing, such as with the SAT, because of the potential for an adverse effect on job opportunities for women, risk of injury, and the variability of physical requirements across jobs (Messing and Stevenson, 1996; Stevenson et al., 1996). The SAT, in particular, has received criticism: As noted earlier, the GAO (1996) identified problems in how the test is administered, cited findings about changes in SAT scores after recruits underwent basic training, and criticized the lack of up-to-date information about physical requirements across specialties that would help determine appropriate requirements.

Next, we describe our initial attempts in fiscal year 2013 to evaluate the relationships between SAT scores from MEPSs and two important types of job outcomes: job performance and workplace injury. To explore these relationships, we reviewed data from Air Force personnel data systems. To measure job outcomes, we considered Enlisted Performance Ratings (EPRs) for job performance and injury rates (with associated information) to measure workplace injury. After thoroughly examining the data, we determined that the assumptions to conduct statistical tests to establish relationships using the selected measures were not met. Consequently, the objectives of this chapter are to describe the challenges and limitations of existing measures to establish the criterion-related validity of the SAT. Relatedly, we also discuss an analysis of test-retest reliability of the SAT, which is a necessary but not sufficient condition for validity.

Challenges in Assessing the SAT-EPR Relationship

Three challenges arise from examining the relationship between SAT and EPR scores. A major challenge concerns the distribution of both sets of scores, which are skewed toward the highest possible values on the respective scales (40 to 110 pounds for SAT, one to five for EPR). In our data set, SAT scores from MEPSs have a mean value of 96.8 pounds, and EPR scores have a mean of 4.6. EPR scores in general and SAT scores for males, in particular, do not vary much, as evidenced by small standard deviations (SDs): SD for total EPR scores is 0.70, and the SD for male SAT scores is 8.2 (with an associated mean of 103.4).[7] Such small variations make it difficult to discern an effect, if one exists at all.

[7] The SD for total SAT scores equals 16.3 pounds. Women's SAT scores have a mean of 70.9 pounds and SD of 14.2 pounds. EPR scores broken out by gender do not differ much: Men's EPR mean equals 4.63 (SD = 0.71) and women's EPR mean equals 4.66 (SD = 0.67).

Another challenge with using EPR and SAT scores concerns changes in specialties and *shreds*[8] over time. As specialties or shreds change (e.g., two specialties merge), two things may occur: (1) sizeable population shifts as people move in and out of specialties or shreds, and (2) changes in SAT requirements. As an example of population shifts, the Operations Intelligence specialty (1N0X1) acquired new members when the Air Force removed the Electronic Systems Security Assessment specialty (1N6X1) in 2009. An example of SAT cut score changes is the change in SAT cut score for Explosive Ordnance Disposal (3E8X1) from 50 pounds to 80 pounds in 2008. SAT scores are reexamined at the request of CFMs and following a standard protocol involving a contractor conducting site visits to three locations for an AFSC to administer short interviews to identify information about the physically demanding tasks performed. An algorithm uses the task information to compute an updated SAT requirement. The estimate produced by the algorithm is then reviewed by the CFM, who may request an adjustment to the SAT requirement to better reflect the physical requirements of the AFSC (Sims et al., 2014). In addition to changes in SAT requirements, changes in specialty population may also affect EPR and SAT score distributions, thus potentially affecting the SAT-EPR relationship.

Another limitation concerns the reliability of SAT scores. Two primary factors affect SAT reliability. The first reflects a general limitation of all physical fitness testing. That is, physical fitness can change relatively quickly. Individuals tested at the MEPS may gain or lose strength by the time they begin BMT. Further changes in strength are also expected as a result of the physical conditioning of BMT. Previous research has shown average muscular strength gains between 4 and 16 percent following BMT (Knapik et al., 1980). After we raised these concerns with researchers at the Air Force Personnel Center (AFPC), two separate analyses were conducted in 2013 to examine test-retest reliability[9] to determine how much change occurs in SAT scores between testing at the MEPS, week-zero at BMT, and week-eight at BMT. The first set of analyses involved using existing data provided by Air Force Recruiting Service to compare scores from MEPS to week-zero at BMT. The second set of analyses involved a new study conducted by AFPC to test the same recruits at week-zero at BMT and again at week-eight at BMT. The same researchers from AFPC administered the SAT at both time points.

Results of SAT score differences between MEPS and week-zero at BMT are shown in Figure 3.1. Out of 61 women and 81 men, less than 30 percent received the same SAT score from the MEPS to BMT week-zero. The majority of score differences were within plus or minus 10 pounds. However, a small percentage of airmen had score differences of 30 pounds or greater. Some of the observed differences indicated airmen received lower scores upon arrival at BMT.

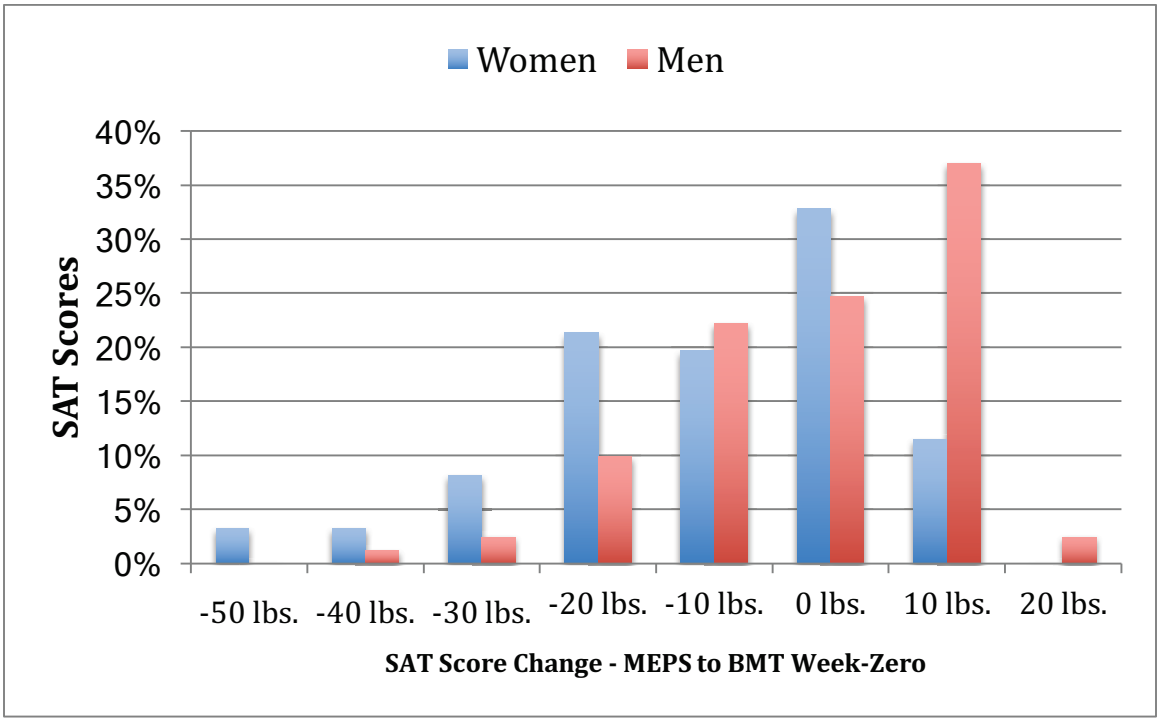
---

[8] *Shreds* or *shredouts* represent subspecialties within an AFSC. For example, Tactical Aircraft Maintenance has three shreds specific to the type of airframe (E=A-10/U-2; L=F-15; M=F-16).

[9] Test-retest reliability is the degree to which participants' scores remain relatively consistent over repeated administrations of the same measure (Crocker and Algina, 1986).

Specifically, 35 percent of men and over 50 percent of women received lower SAT scores. To the extent that these differences are due to actual changes in strength, some airmen may be assigned to AFSCs for which they are no longer qualified. An equally concerning possibility is that the score differences are due to measurement error that may occur as a result of how the test is administered. For example, previous research observed that sometimes recruits were started at a higher weight because they looked strong (Sims et al., 2014). Starting someone at 70 pounds rather than the 40 pounds specified in the SAT protocol may influence the final SAT score achieved. Whether observed score differences are due to actual changes, variations in test administration, or some other cause has important implications for the ways the Air Force could address these differences. For example, additional training and monitoring of test administrators may help to minimize variations in test administration, whereas additional physical fitness training and testing may be needed to address loss of fitness between MEPS testing and BMT. These implications are further discussed in the concluding chapter of this report.

Figure 3.1. SAT Score Differences from MEPS to BMT Week-Zero

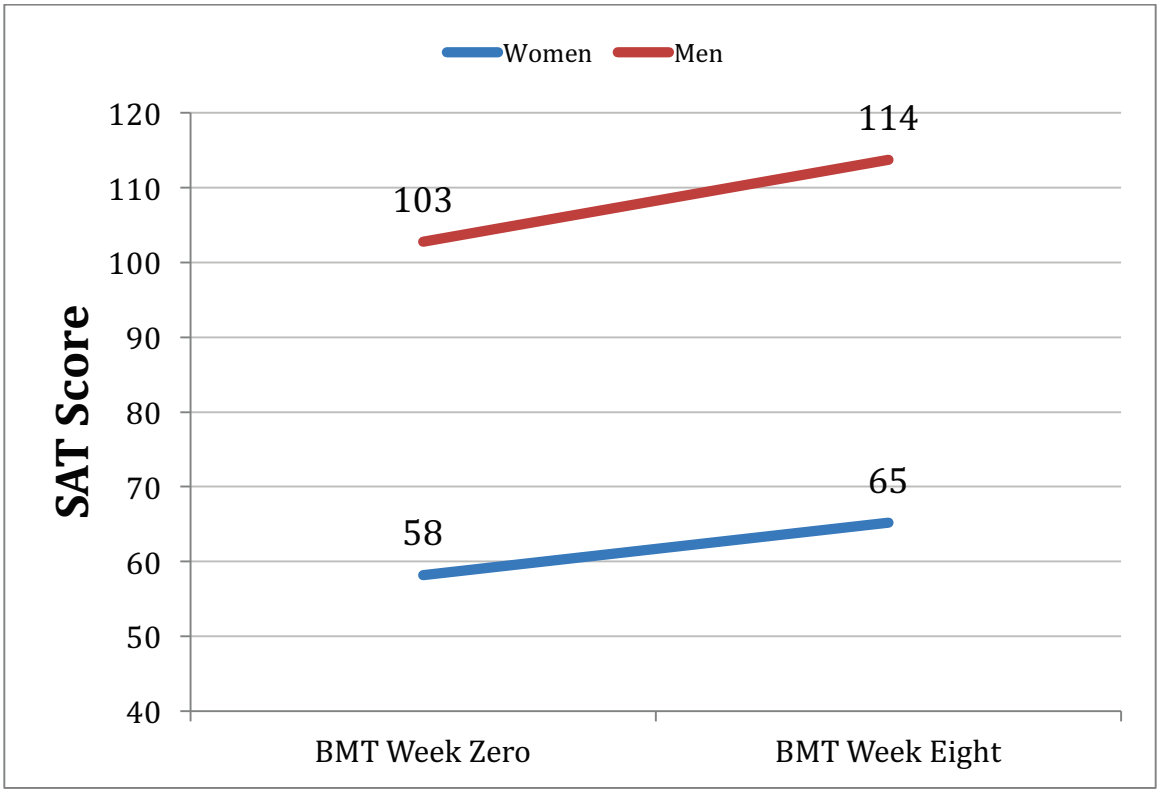


SOURCE: Data provided by AFPC/DSYX (Strategic Research and Assessment).

To deal with the question of potential strength changes that may occur as a result of training during BMT, the Air Force retested a sample of airmen (90 women and 83 men) at BMT week-zero and again at BMT week-eight. To ensure a better estimate of strength gains, the AF allowed

airmen to test up to the full capacity of the ILM, which is 180 pounds.[10] As shown in Figure 3.2, men and women both increased their SAT scores on average. The majority (58 percent) increased their scores by either 10 or 20 pounds, which could be the difference between qualifying and not qualifying for an AFSC. However, a substantial percentage (36 percent) did not increase their scores at all.

Figure 3.2. Average SAT Score Differences from BMT Week-Zero to BMT Week-Eight



SOURCE: Data provided by AFPC/DSYX.

SAT-Injury Relationship

When asked about potential challenges for removing SAT cut score minimums, the most frequent challenge cited by CFMs was avoidance of injury risk (22 percent). Injury risk is an important criterion for selection measures such as the SAT. As summarized by Blakley et al. (1994), Gebhardt and Baker (2010a), and others, personnel selection tests for physical abilities (and requiring a minimum level of physical strength) can ensure that employees in physically demanding jobs can complete the requisite tasks safely; that is, they are less likely to be injured. Note the caveat that this does not apply to all jobs—only to physically demanding ones, which

[10] The maximum score that can be obtained on the SAT at the MEPS is 110 pounds.

typically include firefighters, police officers, and the uniformed services (e.g., combat operatives and those who manipulate heavy equipment as part of their job). Injuries on the job can be costly for organizations and encompass not only medical care but also time lost from work itself as the injured worker recovers (Gebhardt and Baker, 2010a).

We received injury data from the Air Force Safety Center on 216,202 airmen, covering fiscal years (FYs) 2003–2012 and generated from a reporting system designed for preventing lost duty days (Copley et al., 2010). To be reported in the system, the injury must be unintentional and result in one or more days away from work (Copley et al., 2010). Injuries were a very low base rate event; despite the large number of personnel included in the data, only 340 injuries were recorded (N=338 injured individuals, because two people were injured more than once). Descriptions of these injuries ranged from torn muscles during Crossfit training to strained backs while pulling an aircraft fuel hose. We were most interested in work-related injuries, and hence excluded injuries incurred during recreation and leisure activities. This left us with a total incidence of 322 work-related injuries recorded in the data, grouped into the following categories: aircraft ground operations (e.g., "worker strained back while removing aircraft part"), combat support and training (e.g., "injured arm while carrying ruck pack"), government motor vehicle (e.g., "pushing vehicle off railroad track"), industrial and occupations (e.g., "manually opening hangar door; strained lower back"), and miscellaneous (e.g., "worker injured shoulder lifting aircraft part"). All of these injuries occurred while the airmen involved were on duty.

These 322 injuries spread across 71 AFSCs, with 41 percent of the injuries concentrated within only six AFSCs, shown in Table 3.1 along with the proportion of total injuries. Given the low base rate of injuries observed in the data, comparisons by gender and additional statistical tests examining the correlations between SAT requirements and injury rates would not be interpretable.

Table 3.1. Percentage of Injuries by AFSC

Job	SAT Requirement (pounds)	AFSC	Percentage of Injuries
Aerospace maintenance	77.5	2A5X1	9.6
Fire protection	100	3E7X1	8.1
Munitions systems	60	2W0X1	6.8
Aircraft armament	70	2W1X1	5.9
Security forces	70	3P0X1	5.9
Aerospace ground equipment	50	2A6X2	5.3

NOTE: N=44,919; 21 percent of total personnel in data set. AFSC injury variable did not contain shred information, so when a given AFSC had shreds with different requirements, we used the average required minimum score based on the Air Force Enlisted Classification Directory (AFECD) at the time of the study.

The data were limited in the sense that injuries severe enough to be reported to the Safety Center have quite a low base rate in the Air Force. Other less serious types of injuries may not be reported for a variety of reasons, including policy guidance requiring base safety officials to

conduct an investigation for injuries reported to the Air Force Safety Center (Copley et al., 2010), which may act as a disincentive to reporting less serious injuries. Future research should consider whether other sources of data may be needed to accurately measure the full range of possible job-related injuries.

Conclusions

The SAT has not been validated in more than 20 years despite changes to many occupational specialties in the Air Force. This chapter summarized our review of available measures to establish criterion-related validity of the SAT, using existing personnel data to include both personnel performance reports and injury data. Unfortunately, interpretation of any statistical analyses of the relationships between SAT scores and outcomes (i.e., performance evaluations and injuries) was not possible due to limitations in the data. Specifically, each measure suffered from limited variance, which could be caused by a number of factors including poor reliability, actual low incidence of poor performance or injuries, lack of accurate reporting or rating of injuries and performance, and deficiencies in actual measures (e.g., not documenting job-related injuries or measuring job-related physical performance). Based on these considerable limitations, we recommended the Air Force plan and execute a criterion-related validation study using job-related task simulations, which can more directly measure the job-related physical performance of airmen.

Chapter Four. Evaluating the SAT and Related Fitness Tests Using Physical Task Simulations

In 2013, the 1994 Direct Ground Combat Definition and Assignment Rule (DCAR), which excluded women from assignment to units and positions whose primary mission is to engage in direct combat on the ground, was rescinded. As part of the subsequent integration process, Joint Staff guidance and federal laws require that eligibility and occupational standards for all occupations reflect job tasks (DoD, 2013; Pub. L. 113-291, § 524, 2014). To comply with this mandate, in fiscal year 2014, RAND, HumRRO, and the Air Force initiated a study to deal with limitations of previous studies (described in Chapter Three) in examining the validity of the SAT. As part of the study to examine the validity of the SAT, other physical fitness tests were examined to determine whether alternative tests would have stronger validity than the SAT or could be combined with the SAT to improve qualification decisions about the physical capabilities of recruits to perform physically demanding tasks associated with specific AFSCs. This chapter outlines the main questions that needed to be addressed, the overarching methodology to address those questions, and limitations of the selected approach. As described in Chapter One, RAND and HumRRO partnered to execute this phase of the study.

Validation and Study Purpose

Validation involves accumulating relevant evidence to provide a sound scientific basis for how tests, standards, training requirements, and related personnel decisions are applied. The specific type of evidence needed for validation depends largely on the research questions being asked. To address Joint Staff guidance and comply with federal laws, this study was designed to answer the following questions:[11]

1. What are the physical requirements to perform in different AFSCs? (RAND and HumRRO)
2. How can physical performance on job-relevant tasks be measured? (HumRRO)
3. Which physical fitness tests, including the SAT, are valid indicators of a recruit's capability to meet job-relevant physical demands? (RAND)
4. Do the fitness tests predict physical performance equally well for different subgroups (e.g., men and women)? (RAND)

[11] Some questions were addressed jointly by RAND and HumRRO, whereas other questions were addressed primarily by RAND or by HumRRO. The lead is noted for each question in parentheses.

5. How can test scores be used and/or combined to establish qualification standards for current and future AFSCs in the Air Force? (HumRRO)[12]

Based on these questions, a criterion-related validity approach was used for this study. The study used a concurrent (criterion-related) validation design, which involves measuring performance on the fitness tests (i.e., predictors of performance) and the measures of job performance (i.e., the criteria) with the same group of individuals at or around the same time. For this study, the physical performance of Air Force personnel was assessed using physical fitness tests as measures of physical abilities and task simulations as measures of physical job performance.

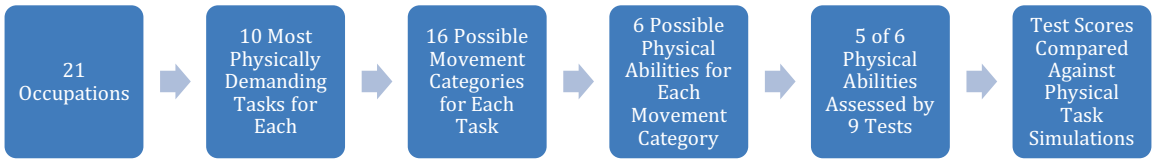
This chapter describes the overall methodology of the study and some important limitations. Although we provide some details on the criterion-related validation study in this chapter, the specific steps and results for each stage of the study are presented in Chapters Five and Six. Appendix J provides further technical details on HumRRO's approach for the criterion-related study and the specific measures used in the study.

Overarching Methodology

The primary steps used to address the first three objectives are presented in Figure 4.1. The Air Force has over 150 different specialties, which prohibits a thorough job analysis of the physical demands for each specialty. Therefore, specific AFSCs (n=21) were sampled to identify physically demanding tasks common to Air Force jobs. RAND and HumRRO collaborated to execute the first task to identify the physical requirements needed to perform in different AFSCs. This task required establishing a job analysis methodology using two primary sources of information, interviews with subject-matter experts (SMEs) and site visits, to systematically evaluate physically demanding job tasks across different AFSCs. This task also involved developing survey items that the Air Force can use to conduct additional job analyses to cover all AFSCs and be able to update standards over time.

[12] As described later in this chapter, attempts to address this question were unsuccessful due to limitations of available data.

Figure 4.1. Steps Completed to Establish Predictive Validity of Tests



Second, HumRRO designed physical performance measures to determine the capabilities of different people to perform physically demanding tasks common to many AFSCs. To develop these measures, physical task simulations were designed to approximate tasks performed across AFSCs. Not all tasks or physical movement patterns were approximated because of limitations in availability of research participants, time, and resources. Consequently, tasks representing the most common movement patterns were used in designing task simulations. Specifically, the task simulations comprised four primary movement categories: (a) lifting and carrying, (b) lifting and holding, (c) climbing, and (d) pushing and pulling. HumRRO established the content validity of the job simulations through a pretest and a reliability study to ensure that the simulation tasks were representative of the physical demands required in each career field. Ultimately, these task simulations assess an individual's ability to perform essential physical tasks across AFSCs and are used to determine which physical tests effectively predict physical job performance.

HumRRO pretested the task simulations on a small group of airmen to determine their viability for use in the validation study. Using airmen in these AFSCs ensured that the task simulations accurately portrayed the physical tasks they perform. The pretest took place at Lackland Air Force base on April 15, 2015. A total of 41 airmen from 18 AFSCs participated in the pretest.[13] The pretest participants were selected to represent each of the 21 sample AFSCs; however, some AFSCs could not participate because of logistical constraints (i.e., availability and proximity to San Antonio, Texas). To ensure an individual's performance on the task simulations would be consistent and reproducible, a reliability study was conducted on May 28–29, 2015, to establish the reliability of the task simulations.[14]

The third step required conducting a criterion-related validity study of physical fitness tests to determine which tests predict performance on physical tasks. The tests considered for evaluation were selected by HumRRO to measure the most important physical abilities required to perform each task. In this study, volunteers from the Air Force[15] completed nine physical

[13] See Appendix B for the email sent to recruit volunteers for the pretest and Appendix C for the numbers of airmen and AFSCs that participated in the pretest.

[14] See Appendix D for the email sent to recruit volunteers for the reliability study.

[15] The validation sample consisted of 412 subjects (278 men, 133 women, one did not report).

fitness tests and four task simulations. RAND then computed statistical models to determine the relationship between test performance and physical task simulation performance. The strength of these relationships formed the basis for establishing evidence for the validity of each physical fitness test. These results further laid a foundation for making recommended changes to physical testing at the MEPSs.

Overall, the study results are designed to generalize to other AFSCs, which require one or more of the primary movement categories: (a) lifting and carrying, (b) lifting and holding, (c) climbing, and (d) pushing and pulling. If the SAT (or another fitness test) is found to predict performance on lifting and carrying tasks sampled from target AFSCs, then the SAT would also be expected to be a good predictor of physical performance for AFSCs that have similar lifting and carrying demands. The specific task being performed is not important, but rather it is the underlying ability required to perform that task that provides the foundation for grouping AFSCs. This logic has been supported by generalizability theory and through job transportability studies (e.g., Hoffman, 1999; Hoffman, Holden, and Gale, 2000; Scherbaum, 2005). That is, results can be generalized to new specialties not included in the study to the extent that they share similar physical demands. Typically, such extensions can be made by comparing the results of the job analyses for the target jobs to job analysis results from the population of jobs (e.g., amount of weight lifted, percentage of assistance with lift). Similar procedures have been successfully implemented in the gas industry to establish validity evidence for physical ability tests (Hoffman, 1999).

Scoping the Next Steps

Ideally, a job analysis would be conducted first for all AFSCs to identify the levels and range of physical demands across the Air Force. Once all job analysis data are collected, the research team would be able to sample AFSCs for the study based on the types and range of physical demands in the Air Force. However, there was not sufficient time to conduct job analyses for all AFSCs prior to evaluating the predictive validity of the SAT. Consequently, RAND and HumRRO decided to sample physical demands from various AFSCs. Without job analysis data on the physical demands of each AFSC, the only data available to benchmark physical demands were the existing SAT requirements (e.g., 40, 50, 60 pounds) and information found in Air Force Occupational Analysis Reports (OARs). Each report is a job inventory of tasks performed by personnel within a specialty, the percentage of time spent performing each task, the number of personnel performing each task, and some summary analyses (e.g., frequency of performance across subgroups such as deployed compared with home station). Although not designed to identify physical demands required to perform job tasks, HumRRO's expertise developed over many years from working with similar occupations provided additional knowledge for reviewing OAR job tasks.

Therefore, HumRRO sampled from the population of jobs using (a) the existing SAT requirement, (b) its subject-matter expertise on the types of tasks performed, (c) career grouping, and (d) review and input from four active and former Air Force officers at RAND. The goal of sampling was to ensure AFSCs represented the most common physical demands required to perform physically demanding job tasks in the Air Force. A full representation of the range of physical demands was not attempted, since tasks requiring an ability that a relative minority of AFSCs perform will have limited utility for recommending a MEPS test given to every Air Force recruit.

Initially, there was some uncertainty in how well the 21 AFSCs in the study represent the general population of AFSCs in the Air Force. Acknowledging that the sample of AFSCs selected for the study may have missed identifying one or more common movement categories, RAND compared the required movement patterns of the sampled AFSCs to the job analysis results from a broader population of AFSCs from the job analysis survey that RAND administered toward the latter stages of the study (described later). This comparison found no additional movement categories to be more common to the broader population of AFSCs than those movement categories represented by the target AFSCs. This finding suggests that the underlying physical abilities required by most physically demanding AFSCs were represented in this study. Although other, less common physical abilities (e.g., anaerobic power used to sprint) may be required for some AFSCs, implementing a test at the MEPS would be inefficient and costly, considering that it would apply only to a small subset of AFSCs. For AFSCs with less common but physically demanding requirements (e.g., battlefield airmen), a more tailored physical ability test would be more effective.

An important limitation of this study is that a direct link between physical test performance and minimally acceptable job performance could be made only in limited cases due to constraints in data available that defined minimum requirements for the AFSCs. HumRRO clustered the AFSCs into groupings by physical demand for each physical ability of interest (e.g., muscular strength). Although providing distinct categories for a physical ability (e.g., low, moderate, high muscular strength) provides a means to classify AFSCs, it did not provide sufficient information on the specific strength required on a test for a group of AFSCs due to the limitations on the validation sample.

This linkage can be accomplished in several ways using SMEs (Cizek, 2012; Truxillo, Donahue, and Sulzer, 1996); however, the strongest link requires minimally acceptable performance levels to be established for each AFSC on the task simulations used in the study. For example, what is the maximum allowable time to lift and carry equipment in the task simulation to be considered a minimally acceptable performer for Explosive Ordnance Disposal (EOD)? Once minimally acceptable performance levels have been set, a corresponding SAT score can be identified to ensure airmen assigned to EOD can perform the lift and carry tasks to an acceptable level. This approach, referred to as *criterion-referenced cutoff scores*, is a useful strategy for determining minimum test scores that minimize the probability of placing

unqualified personnel into jobs that they ultimately cannot perform. In addition to using a criterion-referenced approach, a norm-referenced analysis may also be beneficial (Cascio, Alexander, and Barrett, 1988), which would require an analysis of SAT scores of existing personnel in each AFSC. Examining normative data can ensure that increases in SAT requirements are not set so high as to "disqualify" a disproportionately large group of currently successful airmen.

Because these types of information were not available, this study does not provide the direct linkage necessary to establish a relationship between an SAT standard and minimally acceptable performance in an AFSC. To remedy this limitation, specific implementation steps are discussed in Chapter Seven.

The next chapter focuses on the job analysis methods used by RAND and HumRRO to identify the physical requirements of AFSCs.

Chapter Five. What are the Physical Requirements to Perform in Different AFSCs?

The objectives of this chapter are to provide a detailed summary of the job analysis steps followed by RAND and HumRRO to identify the physical requirements of Air Force occupations. We begin by presenting information about the current distributions of occupations and airmen based on SAT requirements and scores, respectively. This distribution helped inform the sampling approach for selecting occupations to represent the level and range of physical demands across the Air Force. The remaining sections of the chapter provide detailed descriptions of the interviews, focus groups, and surveys used to identify physical requirements of AFSC-specific job tasks.

Distribution of SAT Requirements and Airmen in the Air Force

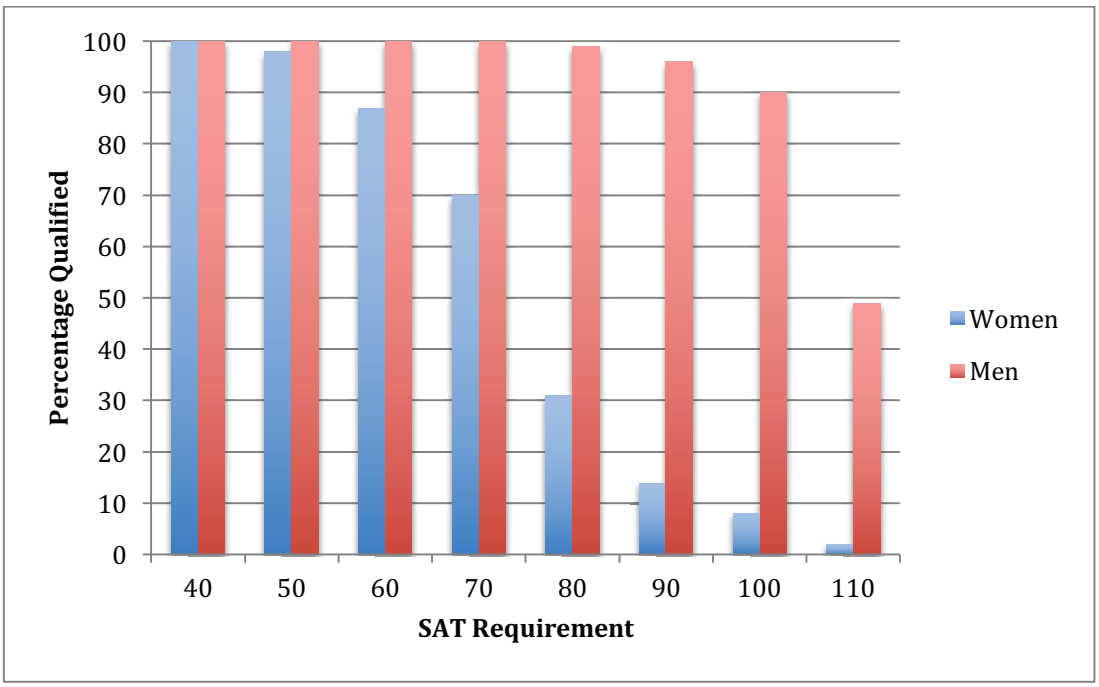
As part of the planning stage of the study, RAND examined the distribution of occupational specialties within the Air Force in addition to how well recruits score on the SAT. As shown in Table 5.1, approximately 38 percent of the occupational specialties have the minimum physical requirement (40 pounds). About 34 percent of the occupational specialties have an SAT requirement higher than 70 pounds (AFECD, 2013). The table also provides information on how men and women are distributed across AFSCs requiring a specific SAT score. Based on data from 2015, approximately 38 percent of the men and 24 percent of the women are working in one of the 33 AFSCs that require an SAT score of 70 pounds. Finally, we present the historical percentages of men and women eligible for AFSCs at each required SAT score based on data from 2000 to 2012.

Table 5.1. 2015 Distribution of Air Force Enlisted Occupational Specialties, by SAT Requirement

SAT Requirement	Number and Percentage of AFSCs	Distribution of Men in AFSCs (%)	Eligibility of Male Recruits 2000–2012 (%)	Distribution of Women in AFSCs (%)	Eligibility of Female Recruits 2000–2012 (%)
40	49 (38%)	22	100	46	100
50	19 (15%)	15	100	16	98
60	16 (13%)	17	100	12	87
70	33 (26%)	38	100	24	70
80	4 (3%)	3	99	1	31
90	3 (2%)	2	96	0	14
100	4 (3%)	4	90	0	7

NOTES: Women are represented in 90- and 100-pound occupational specialties even though the percentage is zero after being rounded. Table excludes most special duty and identifier AFSCs (8x and 9x) and AFSCs with shreds that have different SAT requirements, which includes approximately 14 percent of women and 13 percent of men. AFSCs with shreds that have same SAT requirement were counted as one. Data provided by AFPC and AFRS. Columns may not sum to 100 due to rounding.

Figure 5.1. Historical Qualifying Rates (2000–2012) for Different SAT Requirements



NOTE: As of 2015, no AFSCs require an SAT score higher than 100 pounds, even though recruits are allowed to lift up to the 110-pound maximum at the MEPS.

The historical rates provided in Table 5.1 are extracted and illustrated in Figure 5.1, where it is even more apparent that the SAT begins to have a noticeable influence on the eligibility rates of women for AFSCs with a 60-pound requirement. However, given that women represent an overall small portion of the Air Force and that only about 10 percent of women are screened out

by this requirement, the potential utility of the SAT is minimal. Furthermore, the potential gains in strength from BMT would further limit the potential utility of the SAT for AFSCs below a 70-pound requirement. In contrast, the number of specialties (n=57) requiring 70 pounds combined with further reductions in eligibility rates for women (70 percent) indicate significant potential gains in utility from using the SAT to assign airmen into occupational specialties requiring an SAT score of 70 pounds or greater. For these reasons, RAND initially targeted AFSCs with an existing SAT requirement of 70 pounds or more, which comprises about 34 percent of the AFSCs in the Air Force.

Job Analysis Methodology

RAND conducted the first step of the job analysis to identify the ten most demanding physical tasks for each AFSC in the study, along with ergonomic parameters that defined the physical demand. This job analysis was performed based on two sources of information: RAND and HumRRO interviews with SMEs and HumRRO site visits to Air Force bases to observe task performance. Both relied to some extent on background information found in OARs, which are updated about every three years by analysts from the Air Force Occupational Analysis Division. Although OARs provide important data for classifying, training, and promoting airmen, the tasks listed in OARs do not provide sufficient detail on the physical tasks performed by airmen in each specialty. Consequently, OARs provide limited information on the physical requirements to perform tasks in different Air Force specialties (AFSs). To address these limitations, RAND and HumRRO developed a methodology for identifying job-relevant physical requirements. Described in more detail in the following sections, the methodology integrates scientific and professional best practices for conducting job analyses.

This section examines how information on the physical requirements of each AFSC in the study was gathered through different methods: first interviews, then site visits. As mentioned previously, the study emphasized AFSCs with an existing SAT requirement of 70 pounds or more because this is the point at which the SAT is most likely to begin having utility for the Air Force. Out of 212 AFSCs,[16] 73 have an SAT requirement of 70 pounds or more. Since the job analysis requires identifying the most physically demanding tasks, we initially focused on those AFSCs that had an OAR available. The OAR was seen as an important resource for guiding the identification of tasks that may require physical effort, even though OARs generally do not provide detailed information about the physical tasks (e.g., lifting, pulling, pushing) required by personnel in an AFSC. Furthermore, our job analysis approach encouraged SMEs to create new

[16] *AFSCs* here may mean different shreds for an AFSC. Since different shreds may correspond to different types of equipment that require different types or levels of physical effort, they were treated in the study as separate specialties. This number of AFSCs/shreds is based on AFECD (2013), which was the latest one available before the start of the study in fiscal year 2014.

tasks to represent their AFSC's physical demands if there was not an appropriate task statement in the OAR. An OAR was available for 46 out of the 73 specialties with an SAT requirement of 70 pounds or more.

To this initial selection, RAND added seven career fields with requirements under 70 pounds and two career fields with requirements over 70 pounds but without an OAR, following HumRRO's selection of AFSCs. Adding these career fields increased the range of occupationally relevant physical demands that could be approximated in the validation study (discussed in a later section). Increasing the range minimizes the potential for range restriction by representing jobs that have low to moderate physical demands. At this stage, a job analysis of all of these AFSCs should have been conducted prior to designing the validation study. However, insufficient time was available to complete a job analysis for all these AFSCs; therefore, HumRRO sampled from these AFSCs using the criteria described below. To aid in that selection, RAND provided HumRRO with three types of materials: (1) a brief summary of the AFSC, (2) a slide presentation for the AFSC presenting survey background and summary of results, and (3) a listing of all tasks performed in the AFSC as shown in OAR. These materials provided basic information related to the purpose of the job and tasks performed by different airman ranks.

Using their internal expertise and familiarity with a broad range of demands associated with tasks identified in the OARs, HumRRO sampled from AFSCs that required one or more of the following movement categories: (a) lift, (b) carry, (c) push/pull, (d) climb, (e) walk, (f) stand, (g), hold, (h) shovel/dig, and (i) pound/hammer. Priority was given to AFSCs that required multiple movement categories and were from different AFSC groupings. For example, Aerospace Maintenance contains several AFSCs and was considered as one grouping; therefore, only AFSC Tactical Aircraft Maintenance was selected. Next, AFSCs with multiple subspecialties (i.e., shreds) were reviewed to identify the subspecialty likely to have the greatest physical demands. Selecting the one with the greatest demand could result in establishing requirements set too high, especially if the subspecialty is not representative of the physical demands of corresponding subspecialties. This potential limitation should be further addressed by comparing the physical demands for an AFSC across all subspecialties. If the physical demands are not representative, the Air Force may consider setting different standards for subspecialties. However, this may not be an effective strategy if personnel within an AFSC are expected to be capable of transferring between subspecialties over the course of their careers.

HumRRO's initial selection of AFSCs for the study (see Table 5.2) was submitted for discussion among SMEs at RAND, each of whom had prior service in the Air Force. The participants in the group were asked to review the proposed list of AFSCs and to help ensure that the sampled AFSCs generally represented the range and level of physical demands across occupational specialties in the Air Force. This step was important, since the Air Force has no available database on information on the physical demands of AFSCs.

During this review meeting, the group of RAND experts agreed generally with the list proposed and made the following suggestions:

- adding Security Forces and EOD to the list, because of their potentially unique physical demands compared with other Air Force occupational specialties in this sample
- adding some office jobs with low physical requirements
- ensuring occupational specialties are included that have demands associated with personal protective equipment (PPE)
- taking into account the difference in requirements between wartime and garrison.

Following this review, HumRRO considered RAND's recommendation to include Personnel (3S0X1) but decided not to include it after reviewing information about the job tasks. However, HumRRO made the following two additions to its initial list:

- Security Forces and EOD

The other two comments were addressed by including questions in the interviews and survey that cover use of PPE and the distinction between demands occurring in garrison or in wartime. PPE can increase the physical requirements of a job, especially when the equipment is heavy (e.g., body armor) or can make it more difficult to breathe (e.g., self-contained breathing apparatus).

The final list included 23 AFSCs with varying levels of physical demand; however, two AFSCs (1T0X1, 2A5X1B) were eliminated from further analysis because data could not be obtained from interviews and site visits. These subtractions resulted in 21 AFSCs being included in the study. The primary concern with sampling AFSCs at this stage is the risk that certain physical demands will not be adequately represented in the study. As described later in this report, we were able to address this concern following collection of job analysis data from a broader population of AFSCs. Table 5.2 lists the AFSCs selected by HumRRO after the review process.

Table 5.2. AFSCs Selected for the Study

Final List (after review)		
Job #	**AFSC**	**AFSC Title**
1	1A0X1	In-Flight Refueling
2	1A2X1	Aircraft Loadmaster
3	1T0X1*	Survival, Evasion, Resistance, and Escape*
4	2A3X3L	Tactical Aircraft Maintenance
5	2A5X1B*	Airlift/Special Mission Aircraft Maintenance*
6	2A5X2	Helicopter/Tiltrotor Aircraft Maintenance
7	2A6X1	Aerospace Propulsion
8	2A6X2	Aerospace Ground Equipment
9	2A6X3	Aircrew Egress Systems
10	2A7X1	Aircraft Metals Technology
11	2F0X1	Fuels
12	2M0X2	Missile and Space Systems Maintenance
13	2S0X1	Material Management
14	2W0X1	Munitions Systems
15	2W1X1E	Aircraft Armament Systems
16	3D1X7	Cable and Antenna Systems
17	3E1X1	Heating, Ventilation, Air Conditioning, and Refrigeration
18	3E2X1	Pavements and Construction Equipment
19	3E4X1	Water and Fuel Systems Maintenance
20	3E7X1	Fire Protection
21	3E8X1	Explosive Ordnance Disposal
22	3P0X1	Security Forces
23	4B0X1	Bioenvironmental Engineering

* Subsequently removed from list.

Table 5.3 provides the final list of AFSCs by SAT requirement and the corresponding movement categories, identified by the job analysis, that are required to perform physically demanding tasks within each AFSC. Many of the movement categories were required by all of

the final 21 AFSCs included in the study. Only swimming,[17] digging, shoveling, and running were not identified as common physical requirements across the AFSCs in the sample.

Table 5.3. Movement Categories Required by Final 21 AFSCs Included in the Study

	SAT Requirement					
	50	60	70	80	90	100
	2A6X2 2A7X1	2A6X1 2S0X1 2W0X1 3E4X1	1A0X1 1A2X1 2A3X3L 2F0X1 2W1X1E 3P0X1	3D1X7 3E8X1 4B0X1	2M0X2 3E1X1	2A5X2 2A6X3 3E2X1 3E7X1
Lift	X	X	X	X	X	X
Carry	X	X	X	X	X	X
Push/pull	X	X	X	X	X	X
Climb	X	X	X	X	X	X
Stand	X	X	X	X	X	X
Nonstand (e.g., kneel)	X	X	X	X	X	X
Walk	X	X	X	X	X	X
Run			X	X		X
Crawl	X	X	X	X	X	X
Hold	X	X	X	X	X	X
Shovel	X	X	X	X		X
Dig		X	X	X		X
Pound	X	X	X	X	X	X
Swim						
Oper. power tools	X	X	X	X	X	X
Oper. nonpower tools	X	X	X	X	X	X

[17] Although swimming was not identified in this study as a physical requirement, some AFSCs including Pararescue, Combat Control Team, and Special Operations Weather Team require swimming to perform some duties. These specialties have additional physical screening requirements beyond the SAT, which are being evaluated as part of separate studies conducted by the Air Force and RAND.

Interviews with Subject-Matter Experts

The purpose of the interviews with SMEs was to collect preliminary data for identifying the physical demands for occupational specialties in the Air Force. RAND and HumRRO conducted interviews of the occupational specialties selected by HumRRO to provide the data they needed to design the criterion-related validation study. RAND also conducted a separate set of interviews and surveys for additional occupational specialties not selected by HumRRO for the purpose of future classification of each AFSC based on its physical demands.[18]

The interview protocol consisted of the following three steps:

1. identifying SMEs to participate in the study
2. asking SMEs to fill out a Physical Task Matrix identifying the ten[19] most physically demanding tasks from their OARs as well as the ergonomic categories (e.g., lift, carry, run) required for each physically demanding task[20]
3. interviewing each SME about the specific physical demands of the identified tasks.

In the first step, the RAND team contacted the CFMs for the AFSCs identified. These CFMs were informed of the purpose of the study and given an overview of the data-gathering effort. Specifically, CFMs were provided with the following information:

> RAND researchers will be contacting you soon to request your assistance in identifying physically demanding tasks performed by Airmen in the specialties you manage. In the next few weeks, RAND will contact you by email with instructions to review the occupational tasks performed by Airmen in your specialty. To guide your review, RAND will also send the most recent task lists compiled by the Occupational Analysis Division (OAD) for your specialty. As you review the OAD task list, please consider the following:
>
> 1. What are the most physically demanding tasks performed by Airmen in your specialty?
>
> 2. Why are these tasks physically demanding?
>
> 3. Does the physical effort required to perform these tasks vary across duty locations, shreds, or by other factors?
>
> 4. How important are these tasks for achieving overall job/mission performance?
>
> 5. What percentage of Airmen in this specialty is expected to be able to perform these tasks?
>
> If you are unfamiliar with the physical demands of the specialties you manage, please identify alternative SNCOs for RAND to contact. These SNCOs should be familiar with the physically demanding tasks performed by Airmen in the specialty and how these tasks are performed.

---

[18] Additional information about the surveys can be found later in the report. See also Appendix E for the email that was sent to SMEs to ask them to identify physically demanding tasks.

[19] The original plan was to identify 15 tasks. The decision to reduce to ten tasks is described in a later section.

[20] See Appendix F for a snapshot of the Physical Task Matrix.

> Following your identification of the most physically demanding tasks performed in the specialties you manage, RAND will schedule an interview to ask you more detailed questions about those tasks. The information you provide during the interview will serve as the foundation for updating standards for the occupational specialties you manage. Further, the information you provide will ensure the Air Force is in compliance with Public Law 103-160 by ensuring standards are both gender neutral and occupationally relevant.

CFMs were invited to identify other SMEs for the individual AFSCs being examined, when they could not or would not serve themselves as SMEs.

SME selection followed three criteria. SMEs had to

- have at least two years of experience in the career field
- be familiar with the various job tasks associated with the AFSC
- be able to speak to the physical demands currently required by the AFSC.

The study team made every effort to identify multiple SMEs if it appeared that physical demands varied greatly depending on airframe or other factors. When relevant, different SMEs were identified for the different shreds of a given AFSC.

In a second step, SMEs were asked to identify the ten most physically demanding tasks in their specialty using the OAR task list. The tasks selected

- could not be training tasks
- had to be tasks that most airmen in a given specialty would reasonably be expected to perform.[21]

Airmen were allowed to merge tasks from the OAR if they were performed together or in sequence to accomplish an objective. They could also be merged when there were only minimal differences in equipment or procedure. For example, there are several methods for defueling and fueling aircraft, including single-point and over-the-wing methods. These can be combined and restated as "Defuel or fuel aircraft using single-point, over-the-wing, or other methods." As another example, removing and installing wheel assemblies and tire assemblies are listed as separate tasks in an OAR. Since these tasks are performed as part of a sequence, they can be combined as "remove and install wheel and tire assemblies." SMEs who merged tasks were required to write on the spreadsheet a new task statement that was inclusive of each of the more detailed tasks provided in the OAR.

In addition to listing the ten most physically demanding tasks of their AFSC, SMEs were also asked to assess the level of physical demand required by each task from 1 (extremely high physical demand) to 4 (low physical demand). To record this information, the study team provided SMEs with an Excel spreadsheet ("Physical Task Matrix") with columns listing 16 ergonomic categories: lift, carry, push/pull, climb, nonstanding position (e.g., stoop/squat), stand,

[21] These criteria aimed to eliminate tasks that are not important to the specialty or are only performed by a small subset of airmen within the specialty.

walk, run, crawl, hold, shovel, dig, pound, swim, operate powered hand-held tools, and operate nonpowered hand-held tools.[22] SMEs were asked to enter the ten most demanding tasks in the rows and to identify the physical demands for each task by placing an "X" in the cell(s) to specify whether an ergonomic category applies to that task. For instance, if one of the tasks was "Perform a casualty evacuation," the cells under lift, carry, squat, and walk (all ergonomic categories required by a casualty evacuation) should have an "X." SMEs were requested to return the completed Physical Task Matrices to the study team, which would subsequently schedule an interview to ask more specific questions on the ten tasks selected.

In a third step, the RAND team used the completed Physical Task Matrix to structure the interview with each SME on the physical demands of the SME's AFSC. To record interview data, HumRRO developed an Excel-based data collection instrument called the Movement Classification Questionnaire (MCQ), which was reviewed by the RAND team and revised over several iterations.[23]

The purpose of the MCQ was to collect detailed information on the ten tasks identified by SMEs, based on the ergonomic categories listed in the Physical Task Matrix. The MCQ contains 11 separate sheets for collecting data: one for recording general information about the interview and the SME being interviewed, and ten for recording detailed data on each task. Each sheet lists a series of questions on the physical movements and levels of efforts required for each task, by ergonomic category. An excerpt of questions from the "Lift" classification segment of the MCQ is shown in Table 5.4. Other related information about each task was also recorded, including the following:

- general description of task that includes any relevant subtasks
- information on equipment worn or used while performing the task. Equipment of interest included anything weighing more than 10 pounds and any equipment required for the task that is burdensome or difficult to use.

[22] The Excel spreadsheet contained definitions of these ergonomic categories in the second tab/sheet, labeled "Definitions." See Appendix F for an example of the Physical Task Matrix.

[23] See Appendix G for a snapshot of the MCQ.

Table 5.4. Example of Questions and Information Obtained in the Movement Classification Questionnaire

Item #1: Weight maximum (pounds)
Item #1: Weight minimum (pounds)
Item #1: Percentage lifted with assistance
Item #1: Number lifted at one time
Item #1: Number lifted in task
Item #1: Duration of lift (ONLY IF MULTIPLE LIFTS) (minutes)
Item #1: Height lifted to
Item #1: Height lifted from
Item #1: Size Length (ft)
Item #1: Size Width (ft)
Item #1: Size Height (ft)
Item #1: Number of objects lifted without a 1-minute break

During the pretest of this data collection instrument, RAND and HumRRO found out that the time to gather the data was substantial and that ten tasks adequately represented the physical demand of an AFSC. The initial instruction of collecting information on 15 tasks was therefore lowered to ten. Even with this change, the average interview time was 90 minutes instead of the 45 to 60 minutes foreseen initially.

Observations from Interviews

In addition to the 21 AFSCs selected for the study, RAND also conducted interviews with other AFSCs to further document the physical demands required by AFSCs. In total, the RAND team conducted 51 interviews that covered 51 AFSCs.[24] Five AFSCs initially included in the list of interviews did not receive an interview: The study team found that it already had job information on three battlefield airmen specialties (Tactical Air Control Party [TACP], Combat Control, [CCT] and Pararescue [PJ]) from a separate Air Force-sponsored project, which could be used in the present study, and the team was unable to reach the various points of contact (POCs) for Airlift/Special Mission Aircraft Maintenance (C-20) and for Survival, Evasion, Resistance and Escape (SERE).[25] Appendix H provides the list of the AFSCs that were covered by the interviews.

RAND reviewed the data for gaps and inconsistencies that would warrant follow-up or additional interviews. Following this review, the interview data was forwarded to HumRRO for

[24] Some interviews included more than one respondent.

[25] SERE is among several AFSCs in the Air Force that require an additional physical ability screening test, which includes a swim, a run, pull-ups, sit-ups, and push-ups. Therefore, these AFSCs, which include CCT, EOD, PJ, SERE, Special Operations Weather Technician (SOWT), and TACP rely much more heavily on these tests to determine whether someone is physically qualified to perform tasks associated with each of these AFSCs.

inclusion in its analyses of AFSC physical demands. The RAND team drew the following observations from interviews:

- SMEs often identified many individual positions for each task, but during the interviews they focused on the elements of the task that had the largest effect on the physical demands.
- Because physical demands for a particular task could vary greatly depending on the context, it was most useful to focus on the circumstances under which the task was most difficult and then indicate how frequently this was the case.
- The interviews often did not follow the strictly sequential format embodied in the MCQ, so it was important for interviewers to ask clarifying questions and to take additional notes.
- There was often confusion as to the definitions of the positions and which positions applied to certain actions. For example, if a SME indicated that a carry was involved, they frequently also listed a hold even though the holding element only occurred in the course of the carry. Therefore, it was important for interviewers to work with the SMEs to understand better the nature of the tasks they identified.
- The MCQ may need to be adapted to accommodate information on dimensions for objects not easily described in terms of length, width, and height. Also, at times there was difficulty capturing the details of positions such as push/pull when it was not the act of pushing something laterally, but rather pulling something up using a rope or pushing on a large handle or lever.

Site Visits

The study team identified locations for site visits in two stages. First, the RAND team identified Air Force bases that had airmen in the 21 AFSCs under study and compiled a list that was sent to HumRRO. In a second stage, the HumRRO team built a schedule that optimized visits by selecting Air Force bases that had multiple AFSCs.

RAND contacted CFMs for each of the AFSCs selected. The CFMs were asked to identify POCs in the various locations selected. For instance, the CFM for Fuels (AFSC 2F0X1) was asked to help the study team identify a primary POC for the Fuels specialty at Seymour Johnson Air Force Base (AFB) and Fort Belvoir.

HumRRO scheduled a total of 37 site visits, including observations and interviews with personnel from each of the 21 AFSCs at multiple Air Force bases. Air Force bases visited include Andrews AFB (Maryland), Seymour Johnson AFB (North Carolina), Pope Field (North Carolina), Moody AFB (Georgia), Fort Belvoir (Virginia), and Dover AFB (Delaware).

Survey of Physical Demands

The interviews and site visits provided an important source of data for planning and executing a criterion-related validation study (described later in Chapter Six and Appendix J). However, such an approach to identifying job-related physical requirements is time- and resource-intensive. To ensure the Air Force can continue to update the standards for other

occupational specialties not included in the interviews or site visits, RAND developed and tested survey items that could be integrated into the Occupational Analysis Division's job analysis survey. The following sections describe the development of this web-based survey.

Survey Methodology

Participants. RAND provided a link to a web-based survey and a unique password for accessing the survey. The link and passwords were distributed by our study POCs to CFMs who were instructed to identify a minimum of 15 SMEs from each AFSC that they manage to complete the survey. Service members were sent reminders about the survey periodically over the span of two weeks. A total of 2,052 participants completed part (n = 158) or the entire (n = 1,894) survey. We surveyed 268 AFSCs and AFSC shreds during our data collection efforts, which occurred between August and October 2015.

Survey. The web-based survey focused on understanding the extent to which Air Force service members used 15 different ergonomic categories when completing tasks for their jobs, where tasks were defined as "groups of activities to achieve a specific job goal and have a clear beginning, middle, and end." We included these 15 ergonomic categories in the survey because they aligned with the categories HumRRO used in earlier interviews. We pilot tested the survey twice with RAND researchers and Air Force leadership to ensure that the survey functioned as intended, particularly given the complexity of the logic used to link the survey items together across the survey.

The survey consisted of two parts: screener items and detailed items.

Screener Items

The first part contained screener items about the ergonomic categories in relation to physically demanding tasks that participants reported. The goal of these screener items was to determine whether we needed to ask detailed questions about each ergonomic category. The first part of the screener asked participants to list the ten most physically demanding tasks they performed on their job. The participants then selected which of the 15 ergonomic categories applied to the tasks they listed. For the ergonomic categories they selected, they then indicated the extent to which they used the ergonomic categories for those tasks. The ergonomic categories and criteria, including threshold levels for the screener questions, are included in Table 5.5. If the participants indicated that they met the threshold level for any of the ergonomic categories, they were provided more detailed questions about those ergonomic categories in the follow-up portion of the survey.

Table 5.5. Ergonomic Categories and Criteria

Ergonomic Category	Criteria
Lifting	Lifting without assistance equipment, equipment parts, tools, or materials (i.e., boxes, munitions, or plywood) weighing 25 pounds or more at least once per year
Carrying	Carrying without assistance equipment, equipment parts, tools, or materials (i.e., boxes, munitions, or plywood) weighing 25 pounds or more at least once per year
Pushing/pulling	Pushing or pulling wheeled objects (such as carts or handtrucks loaded with equipment parts, tools, or materials) or nonwheeled objects (e.g., furniture, bags, free-standing equipment) without assistance at least once per year or pushing/pulling equipment parts (i.e., HVAC) that are difficult to move twice per week
Climbing	Climbing objects (i.e., poles, electrical tower, ladder) to a height of 20 feet or higher at least once per year
Standing	Standing for at least one continuous hour once per week
Nonstanding	Kneeling, squatting, stooping, or lying down for at least one continuous hour once per week
Walking	Walking for at least ½ mile without stopping for more than one minute
Running	Running for at least ½ mile without stopping for more than one minute
Crawling	Crawling for 20 feet at least once per week
Holding	Holding and maneuvering objects weighing at least 30 pounds while mounting them in equipment without assistance at least once per year
Shoveling	Manually shoveling material
Digging	Manually digging material
Pounding	Manually pounding objects using a heavy tool such as a sledgehammer, pick, or manual tamper at least once per year
Using powered handheld tools	Using handheld drill, hand Sawzall, chain saw, etc. at least once per year
Using nonpowered handheld tools	Using pliers, hammer, ratchet at least once per year

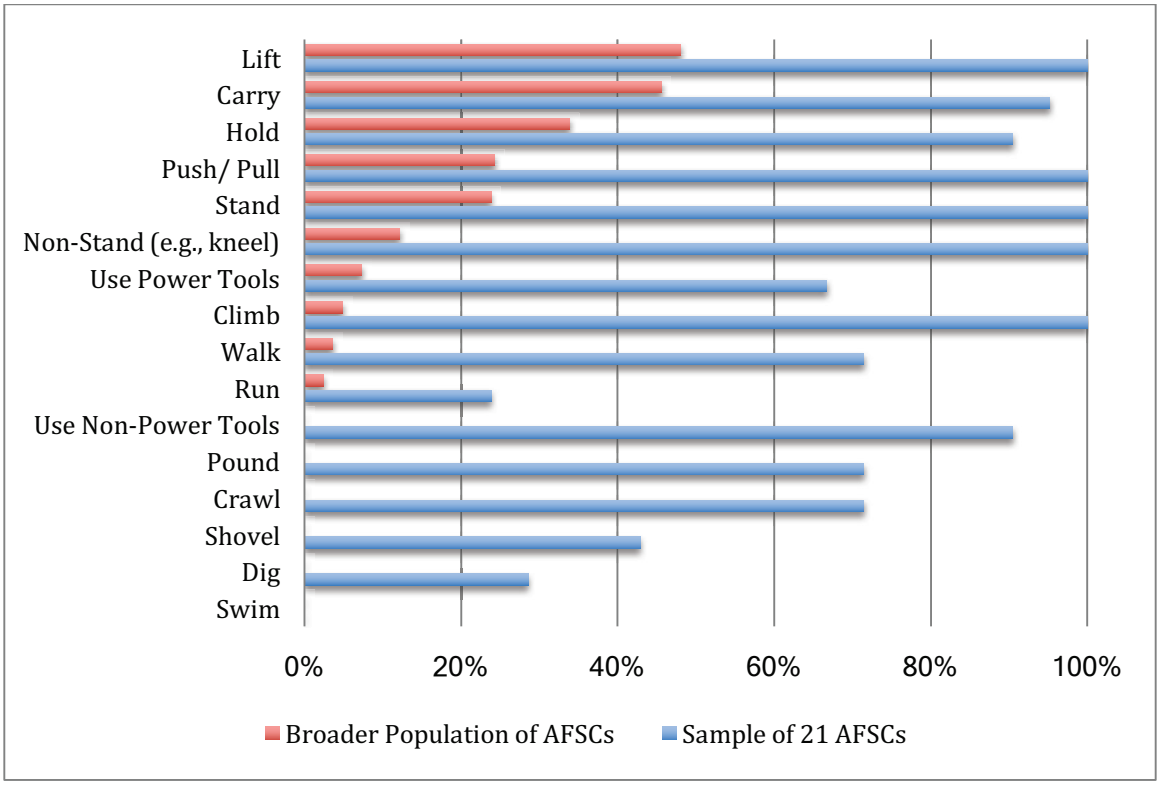
Detailed Items

The second part of the survey contained detailed survey items about the ergonomic categories. Participants were asked to answer these detailed items only if they responded in the affirmative to the screener questions (i.e., participant indicated that he or she lifts at least 25 pounds once per year). As seen in Appendix I, which contains the detailed items by ergonomic category, we asked participants a different number of survey items across these ergonomic categories, ranging from one item for standing to 13 items for carrying. For ergonomic categories where objects are used (i.e., lifting, carrying, holding), the items focused primarily on (1) size (weight, length, width, and height) of objects, (2) ability to use the object without assistance, (3) number of objects needed for a task, and (4) height lifted to/from. For ergonomic categories focused on physical exertion by the participant (i.e., walking, running, crawling), the items asked about (1) pace of the activity, (2) total continuous time spent in this activity, and (3) surface type and slope.

Representativeness of the 21 Selected AFSCs

Some of the limitations previously identified resulted directly from assumptions that had to be made within the limited time available for data analysis following administrative delays in conducting the validation study. For example, HumRRO sampled 21 AFSCs to identify the most relevant movement categories required to perform physically demanding tasks across Air Force jobs. At the time when these AFSCs were sampled, no job analysis information about the physical demands across all AFSCs in the Air Force was available. Therefore, there was no direct approach for determining how well the movement categories represented by the 21 sample AFSCs represent the movement categories of the broader population of AFSCs in the Air Force. To address this concern, RAND conducted a follow-up analysis to compare the movement categories represented by the sample of 21 AFSCs to the broader populations of AFSCs covered in the web-based survey. The web-based survey did not capture physical demand information from all AFSCs, but we received sufficient data from an additional 83 AFSCs, each of which had ten or more respondents complete the survey. In Figure 5.2, the results show the percentage of AFSCs requiring a specific movement category to perform job-related physical tasks (sample of 21 AFSCs in blue; broader population of 83 AFSCs in red). For example, 48 percent of the 83 AFSCs from survey required lifting, whereas all of the AFSCs in the sample required lifting. This difference was expected, since the sampled AFSCs were restricted to those that had lifting requirements.

Figure 5.2. Movement Categories Required by AFSCs in the Study Sample and Across a Broader Population of Jobs in the Air Force



The main finding from these comparisons is that the task simulations in the criterion-related validation study represented almost all of the movement categories required across AFSCs. The two movement categories not represented by the task simulations were the use of power tools and running. Although these movement categories were represented in both the sample of 21 AFSCs and the broader populations of AFSCs, the movement categories were required less frequently than other movement categories and also would have been difficult to simulate in the study because of skill requirements and increased risks to participants. In sum, this analysis increases confidence that the study included the most relevant movement categories required to perform physically demanding tasks in the Air Force. Further comparisons can be made to the full population of AFSCs following completed job analyses for the remaining AFSCs. This step will help eliminate concerns about the representativeness of the physical demands sampled for this study.

Summary

Job analysis is a fundamental step in establishing the physical requirements for an AFSC. We used a combination of job analysis methods including interviews, observations, and surveys to identify the physical demands of AFSCs. At the time of the study, it was not clear which job

analysis variables and movement categories would be most important in identifying the physical demands associated with each AFSC. Therefore, a comprehensive set of questions was initially developed to define the physical demands. If the Air Force integrates physical demand items into its periodic surveys of job requirements, analyses should compare responses across subgroups (e.g., men and women) to ensure results are representative of all subgroups. Furthermore, the Air Force should periodically verify the physical demands being reported by respondents by referencing official documents of equipment that list the dimensions and weights of objects. If these documents are not available, members of the Occupational Analysis Division could conduct site visits and directly observe task performance and weigh the equipment being used.

Chapter Six. Summary of Criterion-Related Validation Study

This chapter provides a summary of RAND's detailed analyses of HumRRO's criterion-related validation study designed to determine the relationships between a range of fitness tests and task performance. A detailed description of the measures used in the study is provided in Appendix J. In the sections following, we evaluate potential combinations of fitness tests designed to meet different objectives and review the extent to which the SAT and a recommended combination of tests equally predict performance for both men and women.

Nine fitness tests were evaluated, including (a) Arm Endurance, (b) Arm Lift, (c) Handgrip, (d) Plank, (e) Push-Ups, (f) Sit-Ups, (g) Standing Broad Jump, (h) Step Test, and (i) the SAT. Task performance was measured using four task simulations designed to approximate physical tasks commonly performed in a range of AFSCs, which included (a) lifting and carrying equipment, (b) pushing and pulling heavy equipment (e.g., tool chest) on wheels, (c) carrying and climbing ladders, and (d) lifting and holding equipment in place. Consistent with HumRRO's analyses, we found that all of the fitness tests significantly correlated with task simulation performance. That is, individuals scoring better on the fitness tests generally performed better on the task simulations.

Evaluating Combinations of Tests

Several methods are available for identifying the combination of tests that best predicts performance. We considered all possible tests (listed in Table 6.1), including the SAT (which is to be used in every model); hence, there are 256 possible subsets of tests to consider. Although comparing this many models manually would be difficult, it is feasible to fit each possible model and compare their respective results using Akaike's information criterion (AIC). For technical information about these analyses, please refer to Appendix K.

Table 6.1 Physical Fitness Tests in the Validation Study

Test	Additional Equipment Cost (yes/no)
SAT	No (already in use)
Arm Endurance	Yes
Arm Lift (mean 3 trials)	Yes
Handgrip—total (mean 3 trials)	Yes
Step Test—VO2 (age-adjusted)	Yes
Push-Ups	No
Sit-Ups	No
Plank test	No
Standing Broad Jump (mean 3 trials)	No

In addition to validity, it is also necessary to consider the resources required for implementing the various test combinations. To account for this trade-off, we prepared a set of options that depend on the balance between the test battery performance and the implementation cost. Specifically, we evaluated the combination of tests that could answer the following questions: Which single test and, likewise, which combination of tests have the highest incremental validity beyond the SAT alone? Similarly, which low-cost test and which combination of low-cost tests have the highest incremental validity beyond the SAT alone? For the sake of comparison, we also consider an option that includes SAT as the sole test. The following is a summary of the options that we examined, and the specific sets of tests that proved optimal for each option are provided in Table 6.2.

- Option 1: SAT is the only test used (baseline)
- Option 2: SAT plus any single test
- Option 3: SAT plus as many other tests as needed
- Option 4: SAT plus any single inexpensive test
- Option 5: SAT plus all inexpensive tests.

When comparing results across all of the options, we recommend the use of Option 2—that is, that only Arm Endurance and the SAT be used if the Air Force has sufficient resources for purchasing arm ergometers for the MEPSs. The use of Arm Endurance provides a sufficient increase in predictive validity, which may justify its cost. Regardless of equipment costs, other options do not provide enough increase in predictive validity to justify the basic resources required for implementation steps such as training test administrators and administering and scoring the tests. While adding simple tests, such as Push-Ups or Sit-Ups, may seem like an easy way to improve recruit screening, the models including these variables performed only marginally better than a model that included only SAT (see Table L.1 for detailed information on these results).

Do the Options Combining Fitness Tests Predict Performance Equally Well for Men and Women?

In addition to predictive performance, another consideration in implementing a test battery is whether it performs equally well for men and women. Gender test bias can occur in several ways and, depending on the nature of the bias, test scores may not be a good indicator of how well a particular subgroup will perform on the job. In the context of physical fitness testing, the presence of test bias could mean a greater proportion of one subgroup (e.g., women) are classified into a specialty for which they cannot perform the physical tasks to an acceptable level. Bias can also result in disproportionately disqualifying more members of a subgroup when they can in fact perform the job tasks.

The following results show that there is statistical evidence for bias in the SAT—that is, the SAT does not perform equally well for men and women. Because the relationship between SAT and physical task simulation performance differs by gender, gender-neutral standards tend to be too conservative for males and too permissive for females. Still, while this bias is enough to be detected with statistical tools, it is unlikely to cause a significant problem in practice for the following reasons:

- The impact of the gender bias depends on the desired level of task performance, and accurate classifications are most critical for physically demanding AFSCs. Despite the tendency of the SAT to overestimate female performance, the necessary SAT thresholds for these AFSCs are high enough that they are unlikely to screen in unqualified female candidates.
- Analysis of the relationship between the SAT and task simulation performance suggest a nonlinear relationship such that gains in SAT scores yield the most gains in task simulation performance at lower levels of strength, and at some point greater strength does not yield better task simulation performance. Accounting for this nonlinear relationship between SAT and key outcomes can mitigate the prediction errors that stem from gender differences.
- Notional screening thresholds for varying job demands indicate that the number of additional classification errors that result from the gender bias in the SAT is likely to be small.

If a more gender-neutral test battery is desired and sufficient resources are available, our analysis also shows that introducing other tests (such as Arm Endurance) can decrease gender bias. The following subsections describe these key points in more detail.

Statistical Tests and Interpretation

We followed the procedures outlined by Lautenschlager and Mendoza (1986) to determine whether tests predicted physical task simulation performance similarly for men and women.[26] That is, for a specific outcome and test battery, we test for the presence of overall gender effects as well as slope effects and intercept effects. A significant slope effect suggests that the relationship mapping a test onto expected task performance differs by gender, and that a test or combination of tests may predict performance better for one subgroup compared with another. If there are no slope effects, but there are significantly different intercepts between groups, this would suggest that performance for one subgroup will consistently be lower than predicted for a range of test scores, whereas performance for the other subgroup will be consistently higher than predicted.

The results indicated that the SAT, when used alone, predicts performance less well for men compared with women. Furthermore, we see that including additional predictors (e.g., Options 2 and 3) reduces gender biases for all outcomes. However, there is still evidence of some gender bias at the 5-percent significance level, even when additional test scores are included in the predictive model. Overall, because of the differences in the distribution of men's and women's SAT scores, no single test or combination of tests fit equally for men and women. Because it is contrary to DoD and Air Force goals to develop different standards for men and women that would predict each gender's performance with equal accuracy, we looked for the single test or set of tests that would minimize the impact of any gender bias.

[26] The specific factors that cause differential prediction are not well known. One possible cause identified in recent research is range restriction (Roth et al., 2014), which can occur in either the predictor or criterion measures. Even though the causes are not well understood, scientific and professional best practices clearly specify the need to evaluate differential prediction to ensure similar test scores have the same meaning for different subgroups (e.g., men and women).

Table 6.2. Differential Prediction Analyses for Gender

		Option 1	Option 2	Option 3	Option 4	Option 5
Physical Fitness Test	SAT	X	X	X	X	X
	Arm endurance	—	X	X	—	—
	Push-Ups	—	—	X	—	X
	Sit-Ups	—	—	—	—	X
	Arm lift	—	—	X	—	—
	Handgrip	—	—	X	—	—
	Plank test	—	—	—	—	X
	Standing broad jump	—	—	—	X	X
	Step test	—	—	—	—	—
R-squared	Climb task simulation	0.262	0.383	0.413	0.280	0.283
	Hold task simulation	0.572	0.597	0.651	0.578	0.615
	Lift and carry task simulation	0.328	0.447	0.483	0.363	0.367
	Push and pull task simulation	0.463	0.545	0.573	0.477	0.479
	Standardized Composite (all task simulations)	0.584	0.703	0.746	0.609	0.617
p-value for gender effects	Climb task simulation	6.18E-04	0.669	0.026	0.027	0.02
	Hold task simulation	0.002	0.327	0.839	0.027	0.291
	Lift and carry task simulation	2.49E-10	0.005	0.022	1.35E-06	8.38E-07
	Push and pull t ask simulation	5.84E-13	3.48E-05	1.21E-03	2.56E-10	3.78E-09
	Standardized Composite (all task simulations)	8.13E-14	1.73E-03	0.009	1.46E-09	3.98E-08

NOTE: Results of tests for slope effects and gender effects are found in Table K.4 in Appendix K.

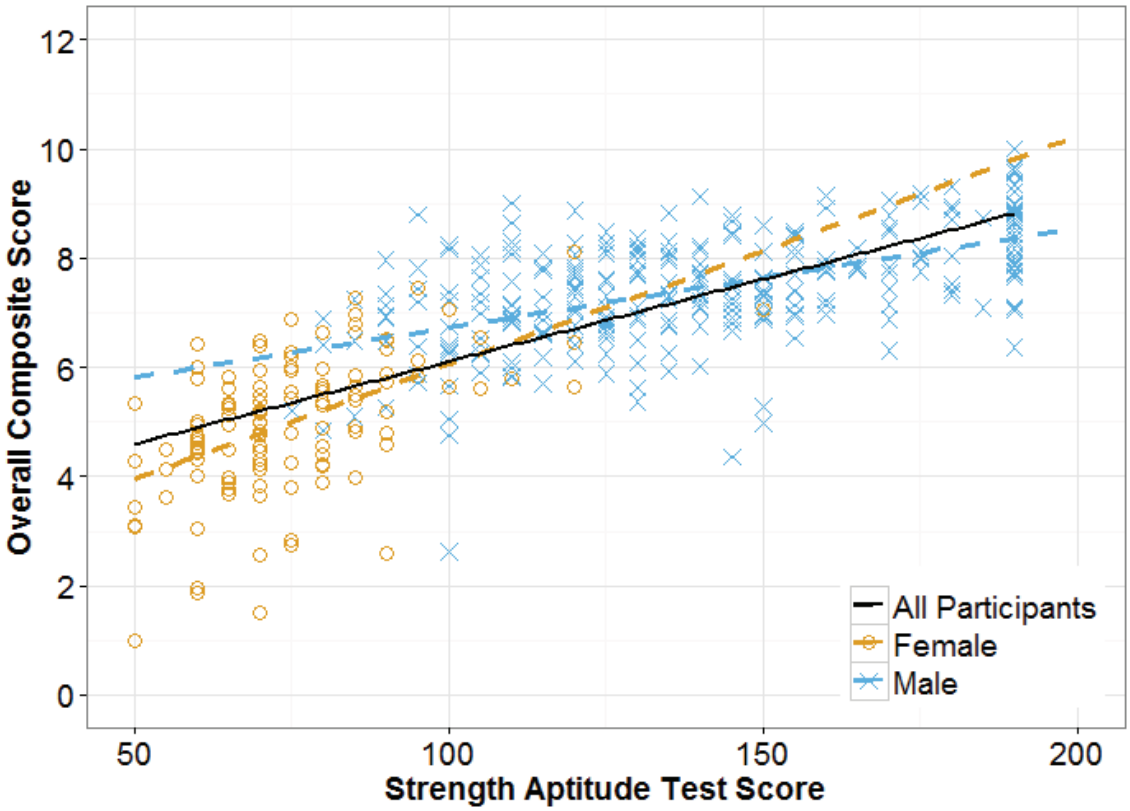
Visual Example Using Linear Regression

To facilitate visualization of the specific gender effects (or how the tests fit men and women differently), we create a series of scatterplots that show all observations in the data while displaying several possible prediction lines using the SAT and Arm Endurance score. When using a model to screen new recruits, the actual task performance is unknown and the predicted performance is used as a proxy. Model performance can be judged by the expected error—the difference between actual and predicted performance. A model with gender bias is one that has systematic differences in the error patterns between men and women.

Figure 6.1 shows a scatterplot of SAT scores against the outcomes composite (created by aggregating standardized scores across all four task simulations) with best fitting lines for all participants combined, as well as for men and for women only. From this figure, there is a clear gender difference in the predicted task simulation score for someone who scores the minimum on

the SAT (the intercept) and in the steepness of the male line compared with the female line (the slope).

Figure 6.1. Predicting Overall Physical Task Simulation Performance for Men and Women Using Only the SAT



NOTE: The model including gender-specific slopes and intercepts has an R^2 of 0.65, versus 0.58 for the linear model with no gender effects. All estimates are statistically significant, with p-values less than 0.001.

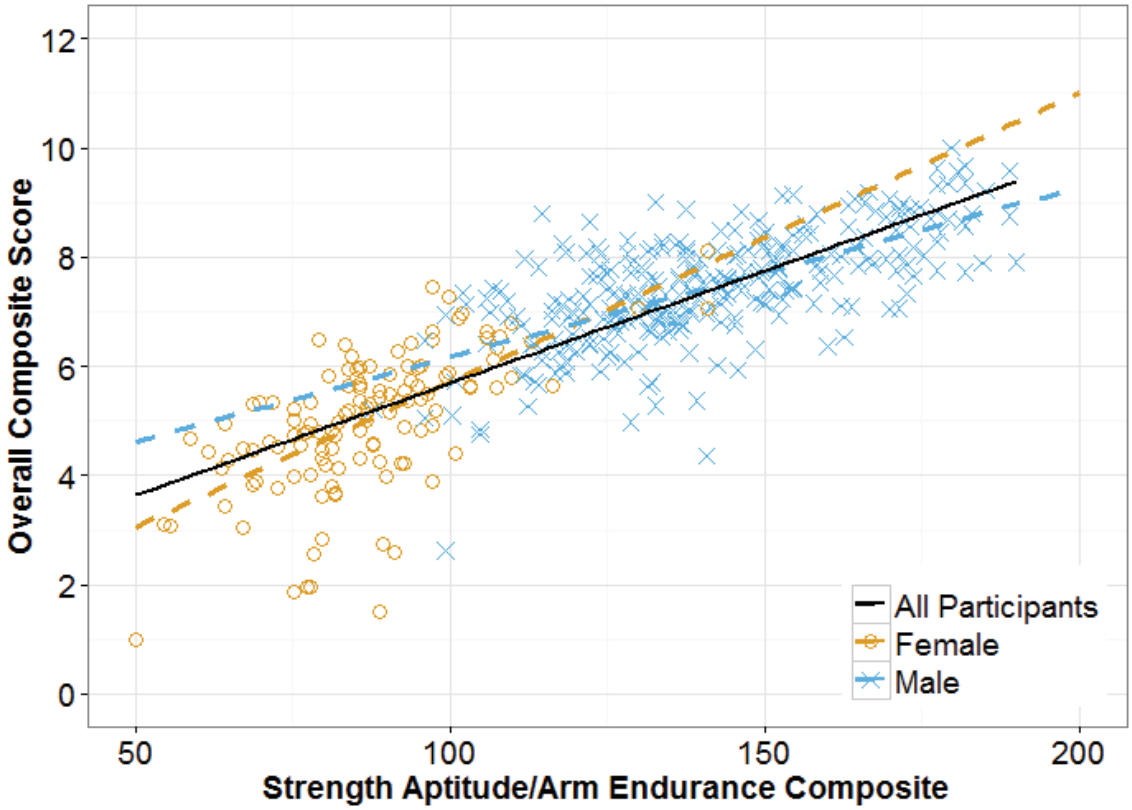
The pattern suggests that the SAT will slightly overpredict performance of women and underpredict performance of men at lower SAT scores. In other words, the risk for the SAT alone is that a few women could be classified into AFSCs for which they may not be able to perform the physical tasks to an acceptable level, and a few men may be excluded from AFSCs for which they would be able to perform at an acceptable level. However, these differences would likely have only a very minor impact on the accuracy of classifying men and women into physically demanding AFSCs. Specifically, the majority of men have historically scored near the maximum on the SAT, so relatively few men would potentially be misclassified. Similarly, the overprediction of women's scores is small at the low end of the SAT distribution and would likely result in relatively few misclassifications.

At the high end of the SAT distribution, the line for women predicts higher performance than the line for all participants. However, very few women in this study scored in this range on the SAT where they would be affected by prediction errors, and the Air Force does not currently have any occupational specialties in the range of SAT scores where this discrepancy occurs. The following section explores the possibility of a nonlinear relationship between SAT and performance as a potential answer to the performance differences among low-SAT versus high-SAT participants.

Combining the SAT and Arm Endurance test with equal weights reduces the gender difference in test predictions (see Figure 6.2). That is, a hypothetical test composite of the SAT and Arm Endurance test yields similar correlations with physical task performance for men and women. However, significant slope and intercept differences remain, so gender bias does not disappear entirely. Inclusion of the less biased Arm Endurance test reduces some of the systematic errors associated with gender, but inspection of the plot reveals that performance for men will still be underpredicted at low levels of test performance, and women's performance will be overpredicted at low levels of test performance. Such a finding presents some amount of increased risk that women could be classified into AFSCs for which they might not be able to perform the physical tasks to an acceptable level. However, the amount of increased risk will depend on the level of minimally acceptable performance for each specialty. Overall, the combination of the SAT and Arm Endurance tests helps to reduce the gender test bias concerns associated with using the SAT as the only physical screening test.

These differential prediction analyses were based on the assumption that each test would be weighted equally in the model. The Air Force may choose to implement these tests using a different strategy (e.g., weighting based on regression weights). Therefore, these analyses may need to be revisited if the Air Force chooses a different system for combining test scores.

Figure 6.2. Predicting Overall Physical Task Simulation Performance for Men and Women Using a Unit-Weighted Composite of the SAT and the Arm Endurance Test



NOTE: To create the SAT/Arm Endurance composite, we calculated z-scores for each test and averaged them, and rescaled the resulting value to match the original range of SAT scores. The model including gender-specific slopes and intercepts has an R^2 of 0.71, versus 0.69 for the linear model with no gender effects. All estimates are statistically significant, with p-values less than 0.001.

Nonlinear Relationships Between Test Scores and Task Performance

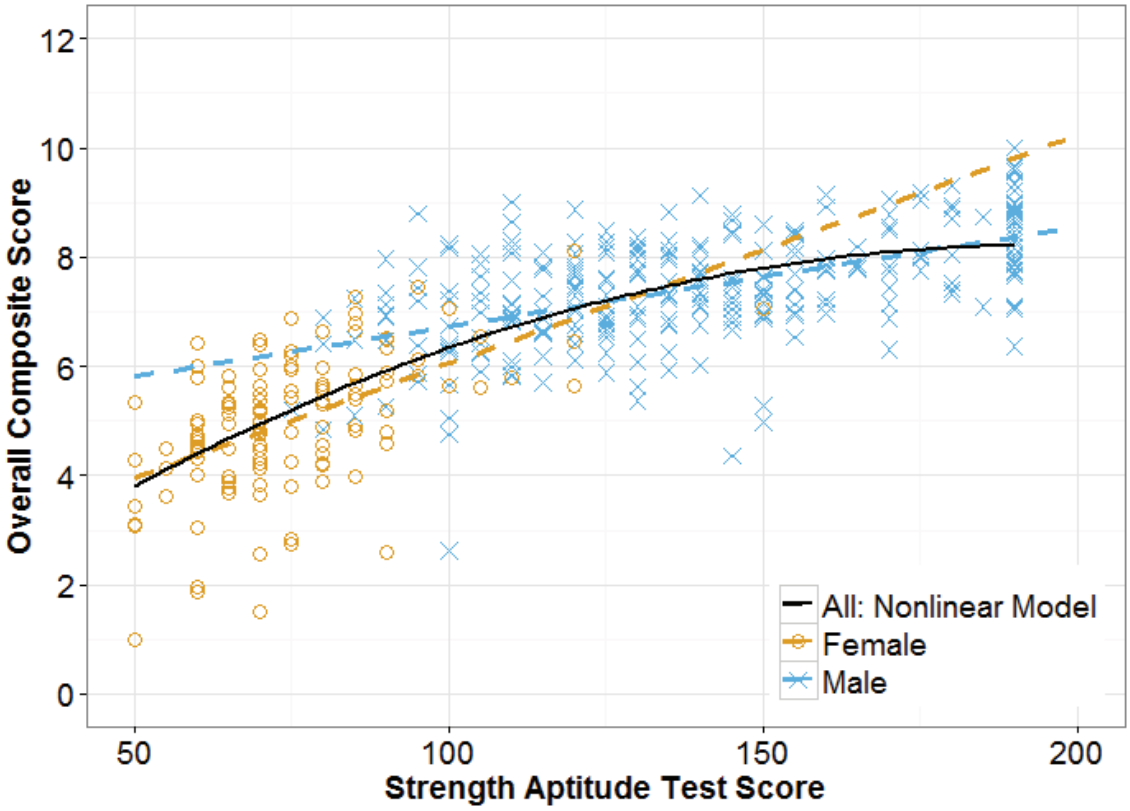
From Figures 6.1 and 6.2, we see that there is little overlap between men and women in either SAT scores or Arm Endurance scores. Assessing gender differences, then, requires the assumption that the steeper relationship observed for women (most of whom scored below 100 on the SAT) would continue for hypothetical women who could score much higher, while assuming the reverse for men—that the flatter relationship among the observed men would hold for hypothetical men who score lower. An alternative explanation for the pattern in Figure 6.1 is that the relationship is nonlinear. A nonlinear relationship might exist if additional strength greatly aided performance for lower ranges of SAT scores, while, at a certain point, further increases in strength were less valuable.

This explanation seems plausible in light of the types of task simulations and level of strength required to perform well in the study. Task simulations were designed to be representative of many Air Force jobs. There is likely a bigger difference in the ability to lift and

carry objects of a standard weight, for example, between those who could lift 50 pounds on the SAT and those who could lift 80 pounds, than there would be between those who could lift 150 pounds and 180 pounds. In other words, the benefits of additional strength could plateau at a certain point, creating a pattern that looks like gender bias but is not.

Figure 6.3 compares this nonlinear alternative with a model that includes a best fitting line for men and women separately, along with the overlaid scatterplot of SAT against the four-task composite score. In this example, the nonlinear model that ignores gender achieves a very similar R^2 value to the model with separate best fitting lines (0.63 vs. 0.65, respectively). The steeper nonlinear curve in the lower SAT range suggests that it may perform similarly well for men and women. The next section will examine this question with a practical example, and additional discussion and analyses on this issue are presented in Appendix K.

Figure 6.3. Predicting Overall Physical Task Simulation Performance for Men and Women Using SAT and Its Square versus Separate Best Fitting Lines



NOTE: The nonlinear model includes only SAT score and its square. The model including gender-specific lines has an R^2 of 0.65, versus 0.63 for the nonlinear model with no gender effects. All estimates are statistically significant, with p-values less than 0.001.

Practical Significance of Potential Gender Bias

Generally, the statistical model with the best fit will tend to minimize screening errors (i.e., cases where a test standard screens out qualified candidates or screens in unqualified candidates). While adding a gender effect to the classification model may be statistically significant based upon the available data, the practical significance for the Air Force (the actual magnitude of the improvement, or reduction in screening errors) may not be large enough to warrant conclusions that a test or test battery is biased and may instead reflect issues with how the tests are implemented (e.g., recruits do not lift to their maximum capacity) or that, in actuality, performances are not distributed along a straight line.

To illustrate the impact of gender bias in the SAT on prediction errors, we created three hypothetical job demand levels and implemented a screening procedure on the population included in the data. While actual task performance would be unknown in a screening situation, it would be possible to administer the SAT (and potentially the Arm Endurance test) and take only candidates with a predicted task performance score at the appropriate level. The illustration assumes that the low-demand category eliminates only candidates who are predicted to fall in the bottom 25 percent of task performance, while the medium-demand category eliminates candidates forecasted to fall in the bottom 50 percent, and the high-demand category accepts only those who the model predicts will perform in the top 25 percent. In this example, we will consider two possible options for predicting task performance: a model based only on screening tests,[27] and a model that manually circumvents potential bias by including gender effects, meaning it uses a different standard to predict the performance of men than it does to predict the performance of women.

The previous scatterplots show that the qualification status of many candidates is unaffected by the screening model. To compare the predictive performance between the models, then, we examine only the cases where the model predictions differ depending on whether gender is considered. The numbers of candidates (of either gender) correctly classified as qualified or unqualified in these "disputed" cases are summarized in Table 6.3. For example, the first row in Table 6.3 indicates that the gender-neutral "SAT Only" model correctly classified 10 qualified candidates for the low-performance categories that the gender-specific model incorrectly deemed unqualified and made one improvement for the medium-performance category and six improvements for the high-performance category. Essentially, the first two rows record a total of 26 "wins" across the three standards for the gender-neutral model over the gender-specific model in the case where the SAT is the only test. The next sets of rows could be considered the "losses" of using the gender-neutral SAT model, as the model that takes gender into account correctly classified two and six qualified candidates for the low- and medium-performance categories that

[27] For models based only on screening tests, we include the test score and its square to account for the nonlinear relationship in the data.

the neutral model incorrectly deemed unqualified and made no improvements over the neutral model on candidates qualified for the high-performance job categories. The third and fourth rows total 28 "losses" for the gender-neutral model, and thus, the gender bias of SAT caused a net increase of two misclassified personnel (out of nearly 1,200 attempts at classifying personnel over the three standards).

Table 6.3. Number of Disputed Observations Correctly Classified, by Hypothetical Job Type and True Qualification Status

		Improvements		
		Job Type		
Tests Used	True Qualification Status	Low	Medium	High
Gender-neutral SAT only	Qualified	10	1	6
	Not qualified	2	6	1
Gender-specific SAT only	Qualified	2	6	0
	Not qualified	10	2	8
Gender-neutral SAT and Arm Endurance	Qualified	0	0	4
	Not qualified	2	3	1
Gender-specific SAT and Arm Endurance	Qualified	1	2	1
	Not qualified	1	0	7
	Total qualified	299	200	100
	Total not qualified	100	199	299

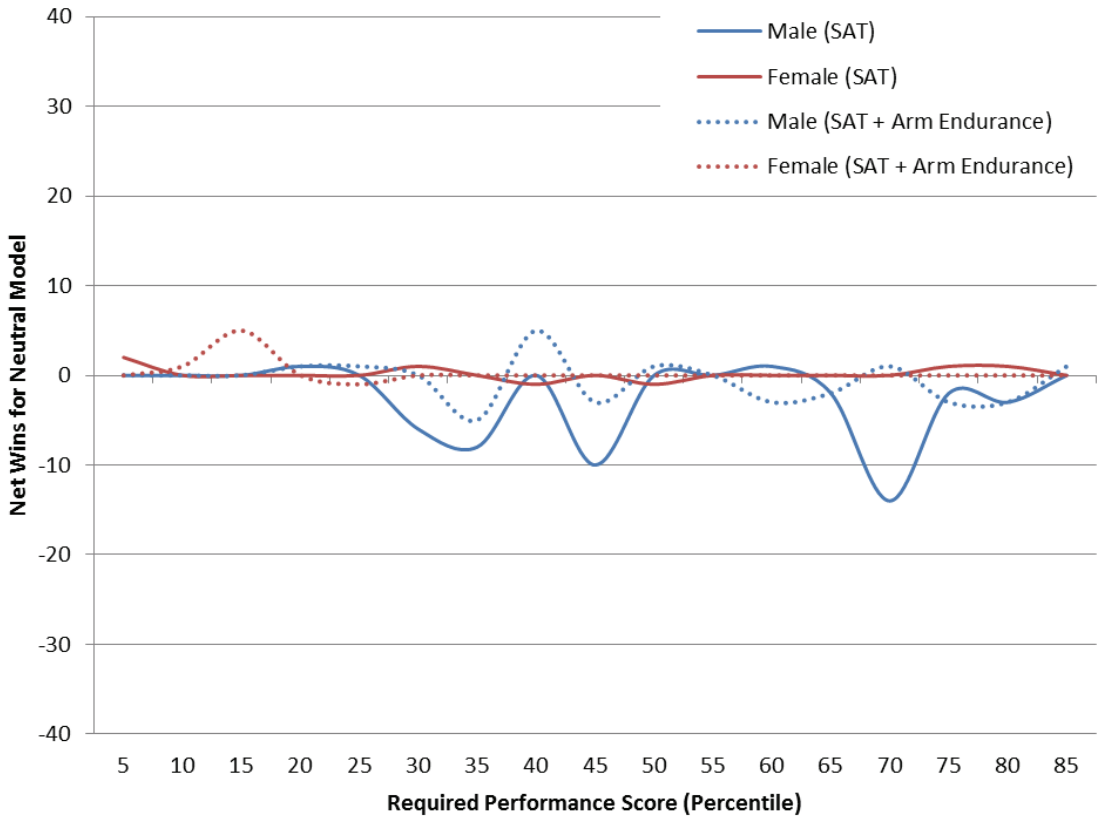
NOTE: Low-, medium-, and high-demand jobs correspond to simulated categories of task performance where workers must perform above the 25th, 50th, and 75th percentiles, respectively. In each case, "disputed" observations are those where models with and without gender disagree. SAT and Arm Endurance indicates a unit-weighted composite of SAT score and Arm Endurance score. Models that include gender estimate a separate intercept and slope for males and females, while gender-neutral models include the respective test score and its square.

The second comparison in Table 6.3, which looks at the SAT and Arm Endurance composite model with and without gender-specific standards, shows results consistent with the earlier finding that using a composite of SAT and Arm Endurance potentially addresses some gender bias concerns. For all hypothetical job types, there are fewer observations where the models disagree, which means that the gender bias has less of an impact on the screening determinations. Still, the model including gender had the same number of net wins across the three hypothetical standards (two), indicating that the end result of the bias is about the same as the model that uses SAT alone to qualify potential recruits.[28]

[28] Net wins were calculated by comparing the sum of improvements for each model. For example, the net wins for the *Gender-neutral SAT only* model were 10 (Qualified-Low), 1 (Qualified-Medium), 6 (Qualified-High), 2 (Not qualified-Low), 6 (Not qualified-Medium), and 1 (Not qualified-High), for a total of 26 net wins.

Finally, Figure 6.4 summarizes the gender bias of the SAT and SAT/Arm Endurance composites by showing the net "wins" (total wins minus total losses for the gender-neutral model) across a broader range of potential job demands, separately for male and female personnel. Figure 6.4 illustrates several realities. First, the impact of gender bias in the SAT is small across the spectrum of potential job demands. Failing to account for gender effects (i.e., using the gender-neutral model) rarely results in more than ten net "losses," which is relatively small as a percentage of the study population. Second, the gender-neutral standard primarily affects the classification of men, who, in this particular sample, show a weaker relationship with the SAT. Lastly, for hypothetical job standards that appear particularly problematic for the SAT (e.g., the 35th, 45th and 70th percentiles), the Arm Endurance test reduces the impact of gender bias and results in fewer net losses for the gender-neutral models. The lines representing the composite of SAT and Arm Endurance (dotted lines) in Figure 6.4 are almost always closer to or above 0 along the distribution of potential required performance levels compared with the models using only the SAT (solid lines), which suggests the composite of SAT and Arm Endurance has a positive influence on reducing gender bias.

Figure 6.4. Net Classification Improvements ("Wins") of Gender-Neutral Model over Gender-Specific Model Versus Required Performance Percentile



NOTE: Values in the figure represent the net improvement (or loss) of a gender-neutral model compared with a model that allows for gender-specific effects. For each required performance percentile, we set the test score cutoff at the score that predicts the corresponding performance composite score. Net "wins" are the number of observations correctly classified by the gender-neutral model that were misclassified by the gender-specific model, minus the number misclassified by the neutral model that were correctly classified by the gender-specific model.

Summary

This study integrated scientific and professional best practices to evaluate the validity evidence for the SAT. Overall, the general pattern of findings supports the predictive validity of all tests considered in this study. Furthermore, we considered how certain combinations of tests could be joined into a battery to achieve different objectives, such as maximizing predictive validity, minimizing cost, and reducing test bias. If any of these additional tests are considered for implementation at the MEPSs, the Air Force should consider further evaluations to balance the relative gains of increased validity with the costs of additional testing and/or equipment.

This study also recommends a way forward for establishing minimum SAT requirements for each AFSC. The results provided by HumRRO should be viewed as a starting point for further

review and analysis during an implementation period. That is, the steps used to generate the job analysis data should be verified with larger sample sizes using the Occupational Analysis Division's routine survey of AFSCs. Furthermore, SAT requirements for each AFSC should be updated as discussed in the next chapter's implementation plan.

Overall, the results from this study support the view that the Air Force has met Joint Staff guidance and federal laws requiring eligibility standards that not only reflect physically demanding job tasks but also are capable of being applied equally to men and women. Although the SAT showed evidence of some differential prediction for women, the potential impact of incorrectly classifying more women or men, on average, would be minimal and does not systematically bias men or women when considering the full range of possible scores on the SAT. In addition to finding strong evidence for the predictive validity of the SAT, we also identified several other tests that could be used in some combination to further strengthen the validity of entry-level physical fitness testing.

Chapter Seven. Courses of Action and Implementation

Based on the program of research reviewed in this report, RAND developed four courses of action (COAs) that the Air Force should consider to ensure enlisted personnel have the physical capabilities to meet the demands of the specialties to which they are assigned. These COAs take into account the body of evidence presented in this report while acknowledging important limitations that must be dealt with during implementation. Each COA can be evaluated on four criteria: validity, cost, ease of implementation, and potential gender test bias. Validity is determined by the average relationship between the test battery and the different outcome variables (i.e., task simulations) from the validation study. Cost can be categorized by the need to purchase additional equipment, maintenance costs, and additional time and resources to administer and score tests. Ease of implementation will generally be influenced by the number of tests and space constraints at the MEPSs. Finally, each COA can be evaluated based on the extent to which test scores associated with the proposed test battery have the same meaning for men as they do for women. In other words, men and women receiving the same test scores would be expected to do equally well when performing physically demanding tasks. Any potential test bias is examined by the presence of significant differences in the slope or intercept, as discussed in Chapter Six. A summary of advantages and disadvantage of each COA and an implementation plan is proposed following a brief overview of each COA.

Courses of Action for the Air Force to Consider

Next, we describe four courses of action with respect to gauging the suitability of a recruit for a range of AFSCs for the Air Force to consider. Each has different strengths and weaknesses, and the choice of one course over the others will depend on validity, cost, ease of implementation, and potential gender test bias. Gender test bias can occur in several ways and, depending on the nature of the bias, test scores may not be a good indicator of a particular subgroup's performance. In the context of physical fitness testing, the presence of test bias could mean a greater proportion of one subgroup (e.g., women) are classified into a specialty for which they cannot perform the physical tasks to an acceptable level. Although there was statistical evidence suggesting gender test bias when using the SAT alone, the magnitude of this effect would likely be small and thus not practically significant for the Air Force. The SAT may err on the side of initially qualifying a higher proportion of women compared with men into AFSCs for which they may not be able to perform the physical tasks to an acceptable level, and excluding proportionally more men compared with women from AFSCs for which they would be able to perform at an acceptable level. Any potential errors in qualification status can be mitigated by conducting job-related assessments in technical training to ensure all trainees are capable of

meeting job requirements prior to being shipped to their first job duty. This recommendation is discussed in more detail later.

For all COAs, the Air Force should maintain the current SAT requirements until additional data can be collected to establish a direct link between SAT scores and minimally acceptable performance in an AFSC. Although HumRRO attempted to form clusters of AFSCs based on shared physical demands, these efforts were only partially successful due to limitations of the validation sample in relation to the AFSCs sampled (n=21), thus resulting in insufficient evidence for updating SAT standards. Therefore, we recommend that technical training courses consider implementing training standards to ensure that all trainees can perform the critical physical tasks associated with their AFSCs. Currently, not all AFSCs with physical demands evaluate trainees on their capability to execute physically demanding tasks during technical training. Trainees are often evaluated on their technical knowledge of how to execute tasks, but not on their physical ability to actually execute the tasks.

This approach provides significant benefits over the more general testing available at the MEPSs. First, technical training can develop task simulations tailored to approximate job-specific, critical physical tasks. The criterion-validation study conducted by HumRRO was well executed given available time and resources but was limited by the number of AFSCs sampled for the study. Even if we were confident that the AFSCs sampled fully represented all physically demanding AFSCs, the task simulations developed for the study had to sacrifice job-related specificity to develop general measures of physical task performance relevant to many AFSCs. Another benefit of implementing physical training standards is that these standards would account for potential changes in physical fitness that result from physical training during BMT and technical training. Therefore, individuals who may not be initially qualified to perform the physical demands of an AFSC have an opportunity to improve their fitness to a level required by the AFSC.

Additionally, individuals will have the opportunity to learn proper techniques to perform such physically demanding tasks as lifting, pushing, and pulling. Implementing a system to measure physical task performance during technical training will provide much-needed data to update the SAT standards at the MEPSs. That is, SAT requirements for each AFSC can be verified against physical task performance during technical training. This process will also provide the most accurate information for establishing SAT requirements that reflect an AFSC's physical demands.

In addition to measuring physical performance during technical training, we strongly recommend that the Occupational Analysis Division integrate physical demand survey items into its regular surveys of AFSCs. The job analysis data collected in this study were based on a small sample of SMEs. Although these SMEs were selected by the Air Force with instructions that they be familiar with the physical demands of their AFSCs, it is possible that the physical demands for an AFSC are not adequately represented, given the range of locations, experience, and diversity of personnel in each AFSC. Collecting survey responses from a broader population

of personnel from each AFSC can provide a more comprehensive understanding of an AFSC's physical demands and further help determine whether different SAT requirements are needed for different subsets of an AFSC (e.g., shreds, assignments, locations).

Another important finding from this research indicated that SAT scores can change over time in response to training at BMT and possibly due to differences in how the test is administered. Therefore, SAT scores from the MEPSs may not be good indicators of an individual's physical readiness to perform physically demanding tasks associated with a specialty. Ideally, the Air Force would test airmen at the end of BMT and use those scores to qualify individuals for different specialties. Although using this approach should provide the best information on an individual's physical readiness, it may not be feasible due to the time required to assign personnel to specialties and training slots. Nonetheless, the Air Force should consider integrating a system of retesting toward the end of BMT for individuals who may not have initially qualified for a specialty in which they are interested or in which the Air Force has a particular need. Such a policy would allow for potential strength gains individuals may have made following the MEPS, either on their own or as a result of BMT. Equally important would be to consider retesting to ensure airmen remain qualified for the AFSCs to which they have been assigned. The Air Force should also consider implementing a system to ensure MEPS test administrators are fully trained and adhere to the SAT testing protocol to promote reliable test administration and scoring. In addition to these recommendations, other factors need to be considered with each of the COAs presented.

COA #1—Adopt the physical test battery at the MEPS that maximizes validity. The combination of tests meeting this objective include the SAT, Arm Endurance, Push-Ups, and Handgrip.

The primary advantage of this COA is that the combination of tests maximizes the potential to ensure recruits have the required physical abilities to perform physically demanding tasks. Furthermore, the combination of tests enables a compensatory model to be developed that more closely approximates how job tasks are performed. More specifically, a compensatory model enables individuals to score somewhat higher on one ability test to compensate for slightly lower scores on another one. For example, an individual with relatively greater muscular strength compared with muscular endurance may be able to perform a job task to a similar performance level as another individual who has relatively higher muscular endurance compared to muscular strength. Although each individual may be able to compensate for lower levels of a specific ability, there are minimum requirements for each relevant physical ability. That is, individuals would not be expected to perform job-related physical tasks unless they met the minimum requirement for each ability, as well as a total combined score across all abilities. A related benefit of this COA is that it provides the most comprehensive assessment of physical fitness to include combinations of tests measuring muscular strength and muscular endurance, which are

required of physically demanding jobs in the Air Force. Analyses also indicated that this test battery resulted in no gender test bias either in slope or intercept differences.

The primary disadvantages of this approach are the increased resource requirements including equipment, personnel, time, and money. Although the Push-Up test does not require additional equipment, significant financial costs are associated with both the Handgrip and Arm Endurance tests, which require purchasing hand dynamometers and arm ergometers, respectively. In addition to the initial purchase cost, additional long-term financial costs are associated with the maintenance of equipment. There may be time and space constraints at the MEPS, thereby increasing the difficulty of implementation and possibly preventing the incorporation of all these tests. Finally, limited evidence is currently available on how best to combine test scores to establish requirements for physically demanding AFSCs.

Although this COA maximizes validity, the return on investment diminishes for each additional test beyond the SAT. Consequently, other COAs may provide alternative strategies for optimally combining tests with minimal validity loss while also minimizing additional costs to the Air Force.

COA #2—Adopt a physical test battery at the MEPS that maximizes validity with no additional equipment costs.

This COA combines the Standing Broad Jump with the SAT. The primary advantages of this COA include increasing validity beyond the SAT with the addition of one test that requires no equipment costs and minimal resources to administer and score. However, the gains in validity (4 percent over SAT) from this COA are minimal and most likely do not justify the additional time and resources required to test and score the Standing Broad Jump, even though there are no additional equipment costs to use this test. Even when all other tests that require no additional equipment costs are added to the model, the validity gained is only 7 percent. Consequently, the validity gains from adding Push-Ups, Sit-Ups, Standing Broad Jump, and the Plank Test are relatively small compared with potential gains from adding tests requiring some additional equipment. Finally, this COA has some evidence of intercept test bias, which will result in overpredicting women's physical task performance and underpredicting men's job-related physical task performance.

COA #3—Adopt a physical test battery at the MEPS that maximizes validity with limited additional costs.

Taking into account the potential advantages and disadvantages of COA #1 and the minimal validity gains from COA #2, our third COA presents an option that balances costs and validity gains. This model combines the SAT with the Arm Endurance Test to provide validity gains on average of 22 percent beyond the SAT alone. These gains are significant and should be considered in light of the potential costs associated with purchasing and maintaining the arm ergometers. Although this COA yields slightly lower validity gains compared to COA #1, it

involves fewer tests, thereby increasing ease of implementation and no costs associated with purchasing and maintaining equipment beyond the SAT and arm ergometer. This model, similar to COA #1, also reduces potential gender test bias.

COA #4—Retain the SAT as the only physical test at the MEPS.

COAs 1 through 3 all require some additional resources to implement at least one or more additional tests at the MEPS. Considering that the SAT has a very strong correlation with physical task performance and that SAT requirements need to be further evaluated and/or updated during an implementation period, the final COA presents a strategy that calls for minimal changes in the near term. Although this COA sacrifices some validity, other factors during the recruiting process likely reduce any losses in utility from using additional tests at the MEPS. For example, self-selection, which involves evaluating perceived fit with job requirements, is one such factor that affects the types of specialties an individual is interested in pursuing. Given a realistic job preview about the physical demands of a job, recruits are likely to pursue jobs that match their perceived qualifications. Individuals who are physically less capable will be less likely to pursue AFSCs that have significant physical demands. The greater the influence of self-selection on the matching process between individuals and AFSCs, the less potential value there will be from using additional physical tests at the MEPS.

Although this COA maintains the use of a valid test at the MEPS, there is some potential concern that the SAT results in gender test bias. Specifically, the SAT, when used alone, does not predict job-related physical task performance equally well for men and women. Even though the magnitude of any potential classification errors is likely to be small as a result of the statistical bias indicated, this COA addresses this concern by further emphasizing the importance of a system to measure the job-related physical capabilities of Air Force personnel during technical training. The additional data from technical training can be used to monitor impacts from any potential gender bias. That is, periodic reviews could evaluate whether the SAT requirements systematically produce disproportionately higher errors for either men or women.

The Air Force should also consider that the majority of AFSCs have few, if any, physical requirements that would benefit from additional physical testing at the MEPS. Therefore, shifting the emphasis of physical testing and standards to these relatively few AFSCs with above-minimal physical requirements may be more appropriate than increasing the number of tests at the MEPS, which would have little, if any, value in classifying a majority of recruits to AFSCs with few physical demands.

Finally, the Air Force should evaluate how well personnel assigned to physically demanding AFSCs maintain their physical readiness. Physical tests such as the SAT can be effective predictors of physical task performance in the short term (e.g., within six months). But physical abilities can change over time; therefore, recruits achieving the required SAT score at the MEPS may or may not be capable of meeting job-related physical demands in the future. The Air Force can mitigate any potential decrease in physical readiness in at least three ways. First, the Air

Force can require annual job-related fitness testing to ensure personnel assigned to physically demanding AFSCs continue to meet the physical test requirements. Second, annual performance evaluations can be redesigned to include supervisor evaluations of performance on job-related critical physical tasks. Third, personnel can be required to demonstrate the physical capability to perform job-related critical physical tasks. This may require developing physical task simulations that can be used as an annual recertification process. Only a few AFSCs may need use this approach or an annual fitness-testing program to ensure physical readiness to perform important tasks that do not occur regularly. For example, AFSCs that require lifting or dragging injured personnel during an emergency may be required to demonstrate physical readiness by either performing the task directly (i.e., physical task simulation) or by demonstrating they have the physical abilities required to perform the task (i.e., annual job-related fitness testing).

This COA does require some additional resources and potential modifications to technical training and may increase the difficulty of implementation. Despite these limitations, significant cost savings can accrue by limiting the number of tests administered at the MEPS, in addition to other benefits from maximizing fairness, ensuring that all trainees are capable of performing job-specific physical demands prior to unit assignment and minimizing any concerns for gender test bias. Overall, the potential gains far outweigh the costs for this COA.

Table 7.1 summarizes the advantages and disadvantages for each of these COAs.

Table 7.1. Advantages and Disadvantages for Each COA

COA	Advantages	Disadvantages
COA #1. Adopt the physical test battery at the MEPS that maximizes validity. The combination of tests that meets this objective includes the SAT, Arm Endurance, Push-Ups, and Handgrip.	• Maximizes potential to ensure recruit has the ability to perform physically demanding tasks • Provides the most comprehensive assessment of physical fitness, to include combinations of tests measuring muscular strength and muscular endurance • No gender test bias indicated	• Requires additional resources and costs for Handgrip and Arm Endurance • May have time and space implications for MEPSs • Return on investment diminishes for each additional test • Evidence on how to combine test scores is limited
COA #2. Adopt a physical test battery at the MEPS that maximizes validity with no additional equipment costs. Combines Standing Broad Jump with SAT.	• Increases validity beyond the SAT with a test that requires no additional costs and minimal resources to administer	• Gains in validity over the SAT (+4%) are minimal and likely do not justify cost and additional resources to administer • Adding in all other no-cost tests still offers limited validity gains over the SAT (+7%) • Test may overpredict female performance and underpredict male performance on tasks (potential gender test bias)
COA #3. Adopt a physical test battery at the MEPS that maximizes validity with limited additional costs. Combines SAT with Arm Endurance test.	• Balances cost and validity gains • Validity increases significantly beyond the SAT (+22%) • Involves fewer tests • Reduces gender test bias compared with using SAT alone	• Slightly less validity gain than COA #1 • Increases costs somewhat for equipment, maintenance
COA #4. Retain the SAT as the only physical test at the MEPS.	• Requires only the SAT test and takes advantage of the relatively strong correlation with physical task performance • Requires minimal changes at MEPSs	• Slightly less validity gain than other COAs • Potential gender test bias

Implementation Plan

Given the study limitations and potential effect on each AFSC, we recommend maintaining the SAT requirements currently in place while following an implementation plan to verify any COA selected by the Air Force. Specifically, we recommend the following steps:

1. Integrate job analysis physical demand survey items into Occupational Analysis Division's routine surveys of each AFSC. The survey items we discuss in Chapter Four can be used. Responses to survey items should be evaluated for differences across subgroups (e.g., location, gender). Periodically verify the accuracy of responses (e.g., weight of equipment) by referencing official documents on the dimensions and weights of equipment, and by directly observing and weighing equipment during site visits.
2. Provide CFMs and other senior leaders in each AFSC with the SAT requirements summary job analysis data for the AFSCs they manage.

3. Collect feedback and address questions or concerns from CFMs and other senior leaders regarding job analysis survey results.
4. Begin administering any new test(s) (e.g., Arm Endurance) at the MEPSs to gather data on new Air Force recruits.
5. Collect data on physical performance of recruits assigned to each AFSC.
6. Use the test data collected from the MEPSs and the physical performance data to verify the accuracy of the SAT requirement and to identify other test scores (i.e., requirements) associated with minimally effective task performance for each AFSC.
7. Calibrate and adjust requirements based on feedback and data collected.
8. Establish system for regular monitoring and updating of test requirements.

First, we recommend that CFMs, Training Pipeline Managers, and Training Cadre review the results from the job analysis survey conducted by RAND to identify critical physical tasks that can serve as a foundation for physical standards in technical training (i.e., used in physical task simulations). Performance on job-related physical tasks can be evaluated if they are important to job or mission performance and any member serving in that specialty would be reasonably expected to be capable of performing the task. Considerations in designing a task simulation include how the task is performed (e.g., one-person compared with two-person lift), the frequency of the demand, task duration, and variability in task requirements by duty location and duty assignment. Task simulations should also be selected to represent the range of physical movements required by each AFSC's critical physical tasks.

In addition to establishing physical task performance standards in technical training, we also recommend implementing a feedback system to monitor whether trainees are meeting these standards. If a certain percentage of trainees (e.g., greater than 5 percent) cannot meet standards, that should trigger a review of the SAT standards for that AFSC. If the SAT requirement is found to be acceptable, an additional physical-demands study conducted by the Air Force Fitness Testing and Standards Unit should be initiated. This study should examine the physical requirements of the AFSC and consider whether additional physical ability screening is required during the recruitment phase. Currently, battlefield airmen and support specialties such as Pararescue, Combat Control, Special Operations Weather Technician, Tactical Air Control Party, EOD, and SERE each have physical screening requirements beyond the SAT.

The SAT has a long history in the Air Force, albeit with concerns raised during its tenure (e.g., GAO, 1996). Physical testing has advantages, particularly for jobs with high physical requirements (Blakley et al., 1994; Gebhardt and Baker, 2010a). The Air Force may wish to consider whether concentration of physical testing resources to the most demanding occupations would enable its most efficient deployment regardless of the COA chosen. As described in this report, only a subset of AFSCs have physical requirements; therefore, focusing efforts on those AFSCs with the greatest physical demands should result in more fidelity and greater efficiency in the overall process. Finally, the COAs described previously all include development of a system to ensure that the Air Force continues to update physical requirements along with changes in the

Air Force jobs themselves, which are key to maintaining the validity of those requirements and, hence, key to ensuring the requirements are beneficial.

Conclusion

This report summarizes a set of studies that addresses previously documented limitations and concerns regarding the SAT. Specifically, we identified that CFMs generally indicated that removing the SAT would present considerable challenges. Therefore, RAND conducted a series of studies and partnered with HumRRO to evaluate the predictive validity of the SAT. RAND's initial attempts to establish a relationship between the SAT, overall job performance, and injuries were unsuccessful in establishing the validity of the SAT. Because of limitations such as range restriction in these outcome variables, we determined that existing measures to establish the predictive validity of the SAT were insufficient. Therefore, we contracted HumRRO to undertake a comprehensive study to develop more-sensitive performance measures using physical task simulations. Using the physical task simulations and criterion-validation data collected by HumRRO, RAND was able to establish that not only can the SAT be used to predict performance on job-related physical tasks but also that other tests (e.g., Arm Endurance) can be combined with the SAT to improve prediction of physical task performance. Now that the predictive validity of the SAT has been established, the Air Force needs to establish minimum SAT requirements for each AFSC. Several courses of action are provided that address this need and will help ensure the physical readiness of recruits assigned to physically demanding AFSCs.

Appendix A: Survey of CFMs on the SAT

Name:

1. Please indicate below which AFS you manage. If you manage **multiple** AFSs **and/or** Shreds, please select and indicate one AFS or AFS-Shred **combination** most familiar **to** you (Please answer the rest of the survey questions based on the AFS/Shred you list here).

2. How familiar are you with the Strength **Aptitude** Test (SAT) – how it is used, **administered,** and **scored?**

Check one.

0 NOT AT ALL FAMILIAR	1 NOT VERY FAMILIAR	2 SOMEWHAT FAMILIAR	3 QUITE FAMILIAR	4 VERY FAMILIAR
☐	☐	☐	☐	☐

3. Please list the current SAT cut score for this **AFS/Shred.**

4. What **physical movements** are required of airmen in this AFS/Shred? How much **physical** effort do these **movements require?**

In the first column, please check all movements that are required. In the second column, please use the dropdown menu to indicate the amount of effort typical of each required movement.

Check all that apply.

PHYSICAL MOVEMENT	REQUIRED	PHYSICAL EFFORT
BALANCING		
BENDING/ROTATING		
CARRYING/LIFTING		
CLIMBING		
CROUCHING/SQUATTING		
GRIPPING/HANDLING		
PULLING		
PUSHING		
RUNNING		
SPRINTING		
STANDING/WALKING		
JUMPING		
REPETITIVE MOVEMENT		
OTHER (PLEASE SPECIFY)		

5. Provide examples of **physical** tasks that airmen would be expected to perform in **this AFS/Shred. Describe tasks that would require the most effort.**
Example: About once a week, airmen are expected to pack cargo bins (weighing about 60-70 lbs.) and place them on pallets without mechanical assistance such as a dolly.

6. Please describe the amount and type of **physical** work (other than **physical training)** required in **technical training** for this AFS/Shred (e.g., extensive **lifting** of boxes over 30 **lbs.).**

7. **Should the SAT cut score for this AFS/Shred be lower, the same, or higher than what it is now?**
Select an option.

If you selected "Same" or "Don't know" above, go to Question 8.
If you selected **"Lower"** or **"Higher"** above, please answer Questions 7a and 7b.

a. What should be the new SAT cut score for this AFS/Shred?
Select Down option.

b. Please describe why you think the score should change.

8. **What are potential challenges or benefits for removing minimum SAT cut score(s) for this AFS/Shred?**
Check all that you believe would be likely to occur.

POTENTIAL CHALLENGES	
INCREASED ATTRITION FROM TECHNICAL TRAINING	
JOB PERFORMANCE WOULD DECREASE	
INCREASED RISK OF INJURIES	
EFFICIENCY WOULD GO DOWN	
OTHER POTENTIAL CHALLENGE(S)? PLEASE DESCRIBE:	
POTENTIAL BENEFITS	
INCREASED MANNING	
INCREASED OPPORTUNITY FOR QUALIFIED MEN & WOMEN TO ENTER AFS/SHRED	
OTHER POTENTIAL BENEFIT(S)? PLEASE DESCRIBE:	

9. **Would you like us to contact you regarding other AFS/Shred's you manage?** If yes, supply your contact information (email and commercial line).

Email:

Commercial Line:

Appendix B: Email Recruiting Volunteers for Pre-Test

Email sent to the CFMs of the 21 AFSCs under study:

[Title, name,]

As you know, the Air Force's Force Management Policy Directorate (AF/A1P) has asked the RAND Corporation to help review and validate the Strength Aptitude Test, which is used to classify airmen into physically demanding AFSCs. We thank you again for your support to our RAND colleague Steve Seabrook during the first phase of this project, which consisted in identifying physical requirements for some of the AFSCs under your purview.

In the next phase of this project, based on the information gathered through interviews and site visits, Human Performance Systems (HPS) has developed a study plan to identify how well the Strength Aptitude Test and other physical fitness tests indicate an airman's ability to perform physically demanding tasks within these specialties. To approximate the tasks performed within these specialties, the Air Force and HPS are constructing four physical task simulations (work samples) to approximate the physical demands associated with these tasks. A brief description of these task simulations is provided below.

To ensure that the tasks reasonably approximate the demands of the following AFSCs: 1A0X1 and 1A2X1, we are asking for your help in identifying airmen in these specialties with a minimum of two years of experience to perform the simulations at Lackland AFB on a date in late March or early April. We need at least one airman from each specialty represented in the study. Our target date is March 25th but there is a possibility for a later date due to potential delays in constructing the simulations. The volunteers will perform the simulations of job tasks (work samples) listed below, which approximate the physical demands and tasks performed in a variety of AFSCs. The airmen will perform a work sample, then discuss it in relation to their job. This process will be completed for each of the four work samples and require about four hours of each airman's time.

Please suggest 2–3 airmen in each of the two AFSCs mentioned (1A0X1 and 1A2X1) who might be able to assist. Also, please let me know if you have any questions about this effort or would like to set up a time to discuss by phone.

Best Regards,

Stephanie Pezard and Sean Robson
RAND Corporation
703.413.1100 x5159

Short description of Task Simulations:

* Lift/Carry

This work sample scenario was designed to simulate the lifting and carrying of objects and equipment in a variety of AFSCs.

* Push/Pull Carts (or AGE)

This work sample scenario was designed to simulate the pushing and pulling of objects and equipment in a variety of AFSCs. The push/pull simulation will involve pushing and pulling carts (or Aerospace Ground Equipment—AGE) of different weights to simulate the differences in cart weights pushed and pulled by airmen.

* Climb Ladders

This work sample scenario was designed to simulate climbing different types of ladders used in a variety of AFSCs. The climbing ladders simulation involves climbing two extension ladders to different heights while wearing a backpack, and then moving objects on a platform.

* Hold Object in Position

This work sample scenario was designed to simulate holding objects in a position during equipment installation or removal procedures found in a variety of AFSCs. The work sample simulates holding objects in a position while the airman holding the objects secures the object in place or holds it while another airman secures the object.

Appendix C: Numbers of Airmen and AFSCs that Participated in the Pre-Test (April 15, 2015)

AFSC	AFSC Title	Number of Participants
1A0X1	In-Flight Refueling	1
1A2X1	Aircraft Loadmaster	2
2A6X1	Aerospace Propulsion	2
2A6X3	Aircrew Egress Systems	2
2A7X1	Aircraft Metals Technology	3
2F0X1	Fuels	3
2M0X2	Missile and Space Systems Maintenance	1
2S0X1	Material Management	2
2W0X1	Munitions Systems	2
2W1X1E	Aircraft Armament Systems	2
3D1X7	Cable and Antenna Systems	2
3E1X1	Heating, Ventilation, Air Conditioning, and Refrigeration	1
3E2X1	Pavements and Construction Equipment	1
3E4X1	Water and Fuel Systems Maintenance	2
3E7X1	Fire Protection	3
3E8X1	Explosive Ordnance Disposal	3
3P0X1	Security Forces	6
4B0X1	Bioenvironmental Engineering	3

Appendix D: Emails Recruiting Volunteers for Reliability and Validation Studies

Email sent to individuals who participated in the pretest:

[Title, Name],

I would like to thank you again for your participation, last April, in the pre-test of the RAND and Human Performance Systems study to help revalidate the strength requirements of physically demanding jobs in the Air Force.

We are now moving to the validation phase of the study, which will determine the tests that are the best indicators of an airman's ability to perform important physically demanding job tasks.

At this time I am asking once again for your help. To conduct the validation study we need 400 volunteers. The study will be conducted at Lackland Air Force Base. Opportunities are available to participate in May and July. Your participation will help to set the standards for the next generation of airmen!

> Participation in May requires about two days of your time from 0700 to 1500 on both May 28th and May 29th.
>
> Participation in July requires about four hours of your time on one day, most likely from 0700 to 1100.

During this time, you will complete several physical fitness tests and four work simulations. More information about these tests is provided in the "Informed Consent" document that will be emailed to you if you are interested in participating.

If you are interested in this study, please send an email to Katie St. Ville at kastville@humanperfsys.com to indicate your interest in participating. Once you send an email to Katie, you will be given a few different documents and further instructions about your participation.

If you have already signed up or contacted Katie, we thank you for your interest and look forward to your participation. We also encourage you to share this opportunity with your fellow airmen. We still need volunteers for next week.

Thank you,

Stephanie Pezard
Sean Robson
(703) 413 1100 x5159

Email sent to CFMs:

Chief [NAME],

I would like to thank you again for your assistance in identifying volunteers a couple of months ago for the pre-test of the RAND and Human Performance Systems study to help revalidate the strength requirements of physically demanding jobs in the Air Force.

We are now moving to the validation phase of the study, which will determine the tests that are the best indicators of an airman's ability to perform important physically demanding job tasks.

At this time I am asking once again for your help. We are looking for 400 volunteers to conduct the next phase of our study at Lackland AFB, which is the validation phase. Some of the testing will take place in July, but our most immediate need is for **additional volunteers for this Thursday and Friday, May 28th and 29th**. I am hoping that, in the event your career field has units/squadrons in the San Antonio area, you could forward them the following information or send us names and contact information of airmen who may be available for the study:

RAND and Human Performance Systems have been asked by senior leaders in the Air Force to help revalidate the strength requirements of physically demanding jobs in the Air Force. Specifically, the Air Force is reviewing the Strength Aptitude Test that measures upper body strength. You may recall taking this test at the MEPS. The Air Force would like to know how effective this test is in determining an airman's ability to perform important physically demanding job tasks, and whether additional tests are needed to ensure airmen can safely and effectively perform their jobs.

Over the past year, we conducted interviews and on-site observations at a variety of bases to determine the physical demands of specialties. Now we are ready to begin the validation study, which will determine the tests that are the best indicators of an airman's ability to perform important physically demanding job tasks.

At this time we need your help. To conduct the validation study we need 400 volunteers. The study will be conducted at Lackland Air Force Base. Opportunities are available to participate in May and July. Your participation will help to set the standards for the next generation of airmen!

Participation in May requires about two days of your time from 0700 to 1500 on both May 28th and May 29th.

Participation in July requires about four hours of your time on one day, most likely from 0700 to 1100.

During this time, you will complete several physical fitness tests and four work simulations. More information about these tests is provided in the "Informed Consent" document that will be emailed to you if you would like more information about the study.

If you are interested in this study, please send an email to Katie St. Ville at kastville@humanperfsys.com and cc smrobson@rand.org indicating your interest in participating. Once you send an email to Katie, you will be given a few different documents and further instructions about your participation.

Thank you,

Stephanie Pezard
Sean Robson

Appendix E: Email sent to Subject-Matter Experts to Identify Physically Demanding Tasks

[Title, name],

Thank you for your participation last year in completing the survey about the Strength Aptitude Test. We are now conducting a much more thorough analysis of the physical ability requirements in physically demanding occupational specialties. We need your assistance to help us document the physical demands of Traffic Management. Below is a more thorough description of the current project.

The Air Force's Force Management Policy Directorate (AF/A1P) asked the RAND Corporation to help identify the physical ability requirements in physically demanding occupational specialties. As part of this effort, we are asking career field managers in these specialties to provide information about the **10 most physically demanding tasks** performed by individuals in the specialties they manage.

This project is approved by AF/A1PT–Chief of Testing Policy and Air Force Examining Activities.

To facilitate the identification of physically demanding tasks, we have attached a copy of the tasks for this specialty copied from the Occupational Analysis Report (OAR) for the specialties you manage. As you review the OAR, please consider tasks that may require one or more of the following ergonomic factors: lift, carry, push/pull, climb, stoop/squat, lie down, kneel, stand, walk, run, crawl, hold, shovel, dig, pound, swim, operate powered hand-held tools, and operate non-powered hand-held tools. The attached Excel spreadsheet contains definitions of these ergonomic categories in the second tab/sheet, labeled "Definitions." When selecting tasks, please do not include physical training tasks among the 10 tasks you list. Further, select only important tasks that most airmen in this specialty would reasonably be expected to perform. That is, please do not select tasks that are not important to the specialty or are only performed by a small subset of airmen within the specialty.

Once you have identified the 10 most physically demanding tasks, please list them on the attached spreadsheet "Physical Task Matrix" in column B. In some cases, the OAR tasks are quite specific and can be merged with other tasks that are performed together or in sequence to accomplish an objective. Similarly, tasks can be merged when there are only minimal

differences in equipment and/or procedure. If it makes sense to merge tasks, please write a new task statement that is inclusive of each of the more detailed tasks provided in the OAR. For example, there are several methods for defueling and fueling aircraft, including single-point and over-the-wing methods. These can be combined and restated as "Defuel or fuel aircraft using single-point, over-the-wing, or other methods." As another example, removing and installing wheel assemblies and tire assemblies are listed as separate tasks in an OAR. Since these tasks are performed as part of a sequence, they can be combined to state "remove and install wheel and tire assemblies."

After listing the 10 tasks, please identify the physical demands for each task by placing an "X" in the cell(s) to specify if an ergonomic category applies to that task. The first row has been completed as an example, which indicates that "Perform a casualty evacuation" requires "lift, carry, squat, and walk."

Once you have completed the attached Excel spreadsheet, please save your responses. Attach and return the spreadsheet by email to Sean Robson at smrobson@rand.org. Once received, we will follow-up to schedule an interview to ask more specific questions about the physical demands in this specialty. Depending on the complexity of the physical demands in the specialty, interviews may take between ***45–60 minutes***.

RAND will use the responses you provide as one source of information to recommend physical ability tests and standards for this specialty. Although we will not connect your name with your responses in our analysis and reports, your unique position as a career field manager means that an informed reader of RAND's report could infer who you are. However, RAND will not ask for sensitive information as part of this study.

If you have any concerns about the purpose of this project, please contact the AF/A1PT POC for operational survey logistics, Mr. Johnny Weissmuller, located in the Strategic Research & Assessment Branch, HQ Air Force Personnel Center:

Phone: COM (210) 565-2238, DSN 665-2238
email: Johnny.Weissmuller@us.af.mil
HQ AF/A1PT–Force Management Policy Directorate
Training & Educational Requirements & Resources Division
Examining Activities Program

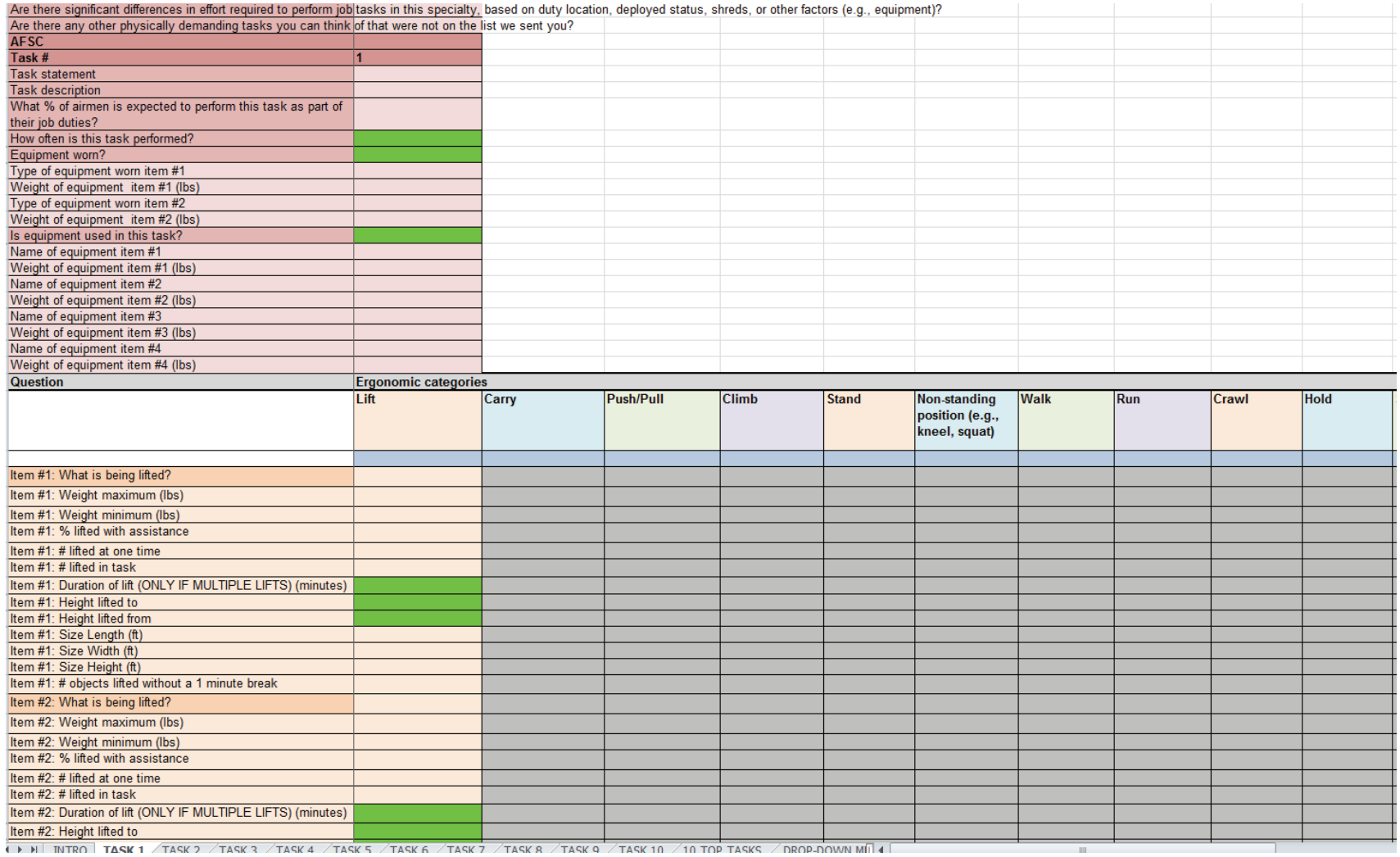
For other questions about this survey, research project, or the RAND Corporation, please contact me using the contact information below.

Sincerely,

Appendix F: Physical Task Matrix

			Ergonomic Categories															
Task #	Physically Demanding Task	Physical Demand	Lift	Carry	Push/Pull	Climb	Non-Standing Position	Stand	Walk	Run	Crawl	Hold	Shovel	Dig	Pound	Swim	Operate Powered Hand-Held Tools	Operate Non-Powered Hand-Held
Example	Perform a casualty evacuation	1= Extremely High Physical	x	x			x		x									
1																		
2																		
3																		
4																		
5																		
6																		
7																		
8																		
9																		
10																		

Appendix G: Movement Classification Questionnaire (MCQ)



Are there significant differences in effort required to perform job	tasks in this specialty, based on duty location, deployed status, shreds, or other factors (e.g., equipment)?									
Are there any other physically demanding tasks you can think	of that were not on the list we sent you?									
AFSC										
Task #	**1**									
Task statement										
Task description										
What % of airmen is expected to perform this task as part of their job duties?										
How often is this task performed?										
Equipment worn?										
Type of equipment worn item #1										
Weight of equipment item #1 (lbs)										
Type of equipment worn item #2										
Weight of equipment item #2 (lbs)										
Is equipment used in this task?										
Name of equipment item #1										
Weight of equipment item #1 (lbs)										
Name of equipment item #2										
Weight of equipment item #2 (lbs)										
Name of equipment item #3										
Weight of equipment item #3 (lbs)										
Name of equipment item #4										
Weight of equipment item #4 (lbs)										
Question	**Ergonomic categories**									
	Lift	**Carry**	**Push/Pull**	**Climb**	**Stand**	**Non-standing position (e.g., kneel, squat)**	**Walk**	**Run**	**Crawl**	**Hold**
Item #1: What is being lifted?										
Item #1: Weight maximum (lbs)										
Item #1: Weight minimum (lbs)										
Item #1: % lifted with assistance										
Item #1: # lifted at one time										
Item #1: # lifted in task										
Item #1: Duration of lift (ONLY IF MULTIPLE LIFTS) (minutes)										
Item #1: Height lifted to										
Item #1: Height lifted from										
Item #1: Size Length (ft)										
Item #1: Size Width (ft)										
Item #1: Size Height (ft)										
Item #1: # objects lifted without a 1 minute break										
Item #2: What is being lifted?										
Item #2: Weight maximum (lbs)										
Item #2: Weight minimum (lbs)										
Item #2: % lifted with assistance										
Item #2: # lifted at one time										
Item #2: # lifted in task										
Item #2: Duration of lift (ONLY IF MULTIPLE LIFTS) (minutes)										
Item #2: Height lifted to										

INTRO | TASK 1 | TASK 2 | TASK 3 | TASK 4 | TASK 5 | TASK 6 | TASK 7 | TASK 8 | TASK 9 | TASK 10 | 10 TOP TASKS | DROP-DOWN MEN

Appendix H: List of AFSCs Interviewed by the RAND Team

Table H.1. AFSCs Interviewed by the RAND Team

SAT Cut Score (pounds)	AFSC/ Shred	Title	Number of SMEs Interviewed
70	1A0X1	In-Flight Refueling	2
70	1A1X1	Flight Engineer	5
70	1A2X1	Aircraft Loadmaster	1
70	1P0X1	Aircrew Flight Equipment	1
70	2A3X3E	Tactical Aircraft Maintenance: A-10/U-2	1
70	2A3X3L	Tactical Aircraft Maintenance: F-15	1
70	2A3X3M	Tactical Aircraft Maintenance: F-16	1
70	2A3X8A	Remotely Piloted Aircraft Maintenance: MQ-1/MQ-9	1
70	2A3X8B	Remotely Piloted Aircraft Maintenance: RQ-4	1
80	2A5X1B	Airlift/Special Mission Aircraft Maintenance: C-130/C-27J	1
80	2A5X1C	Airlift/Special Mission Aircraft Maintenance: C-5	1
70	2A5X1D	Airlift/Special Mission Aircraft Maintenance: C-17	1
100	2A5X2	Helicopter/Tiltrotor Aircraft Maintenance	1
60	2A6X1	Aerospace Propulsion	1
50	2A6X2	Aerospace Ground Equipment	1
100	2A6X3	Aircrew Egress Systems	1
70	2A6X5	Aircraft Hydraulic Systems	1
70	2A6X6	Aircraft Electrical and Environmental Systems	1
50	2A7X1	Aircraft Metals Technology	1
70	2F0X1	Fuels	1
70	2M0X1A	Missile and Space Systems Electronic Maintenance: ICBM	1
70	2M0X1B	Missile and Space Systems Electronic Maintenance: Cruise Missiles	1
90	2M0X2	Missile and Space Systems Maintenance	1
70	2M0X3	Missile and Space Facilities	1
60	2S0X1	Materiel Management	1
70	2T0X1	Traffic Management	1
60	2W0X1	Munitions Systems	1
70	2W1X1C	Aircraft Armament Systems: A10	1
70	2W1X1E	Aircraft Armament Systems: F-15	1
70	2W1X1F	Aircraft Armament Systems: F-16	1
70	2W1X1J	Aircraft Armament Systems: F-35	1
70	2W1X1K	Aircraft Armament Systems: B-52 and B-2	2
70	2W1X1L	Aircraft Armament Systems: B-1	1
70	2W1X1N	Aircraft Armament Systems: F-22	1

SAT Cut Score (pounds)	AFSC/ Shred	Title	Number of SMEs Interviewed
70	2W1X1Z	Aircraft Armament Systems: All Other	1
70	3D1X3	RF Transmission Systems	4
80	3D1X7	Cable and Antenna Systems	3
90	3E0X1	Electrical Systems	1
70	3E0X2	Electrical Power Production	1
90	3E1X1	Heating, Ventilation, Air Conditioning, and Refrigeration	1
100	3E2X1	Pavements and Construction Equipment	1
70	3E3X1	Structural	1
60	3E4X1	Water and Fuel Systems Maintenance	1
100	3E7X1	Fire Protection	1
80	3E8X1	Explosive Ordnance Disposal (EOD)	1
70	3P0X1	Security Forces	1
80	4B0X1	Bioenvironmental Engineering (BE)	2
70	8M000	Postal	1

Appendix I: Survey Items

Table I.1. Survey Items

Ergonomic Category	Detailed Item/*Instructions*	Response Option
Lifting		
	In this task, what objects must be lifted that weigh at least 25 pounds? Please list up to three.	Open-ended text (50 characters maximum; 3 response options)
	Instructions: Please provide the following information for the [first/second/third] object you entered:	
	Minimum weight	____ pounds (range = 1 – 999)
	Maximum weight	____ pounds (range = 1 – 999)
	Length	__ feet (open-ended number)
	Width	__ feet (open-ended number)
	Height	__ feet (open-ended number)
	Percentage of time someone else helps with lifting	_____ % (range = 0 – 100)
	If percentage above is greater than zero: How many other people assist?	___ (range = 1 – 999)
	Number lifted at any one time	___ (range = 1 – 999)
	Number lifted to complete the task	___ (range = 1 – 9999)
	Height lifted to	1 = above shoulder level; 2 = shoulder level; 3 = chest level; 4 = waist level; 5 = knee level; 6 = ankle or ground level; 7 = other
	Height lifted from	1 = above shoulder level; 2 = shoulder level; 3 = chest level; 4 = waist level; 5 = knee level; 6 = ankle or ground level; 7 = other
Carrying		
	In this task, what are the objects that must be carried that weigh at least 25 pounds? Please list up to two objects.	Open-ended text (50 characters maximum; 2 response options)
	Instructions: For each object that is carried, please provide the following information:	
	Minimum weight	____ pounds (range = 1 – 999)
	Maximum weight	____ pounds (range = 1 – 999)

Ergonomic Category	Detailed Item/*Instructions*	Response Option
	Length	__ feet (open-ended number)
	Width	__ feet (open-ended number)
	Height	__ feet (open-ended number)
	Percentage of time carried with assistance	_____ % (range = 0 – 100)
	If percentage above is greater than zero: How many other people assist?	___ (range = 1 – 999)
	Number carried at any one time	___ (range = 1 – 999)
	Number carried to complete the task	___ (range = 1 – 9999)
	Duration of carrying	____ minutes (open-ended number; range = 1 – 999)
	Distance carried	1 = 25 feet or less; 2 = 26 to 50 feet; 3 = 51 to 100 feet; 4 = 101 to 200 feet; 5 = ¼ mile; 6 = ½ mile; 7 = 1 mile or more
	Number of stairs walked up while carrying	0 = I do not walk up stairs while carrying; 1 = 1 to 5 stairs; 2 = 6 to 10 stairs; 3 = 11 to 18 stairs or one floor; 4 = 19 to 35 stairs or two floors; 5 = 36 or more stairs or 3 or more floors
Pushing/pulling wheeled object		
	What is the wheeled object that is being pushed or pulled?	Open-ended text (50 characters maximum)
	Minimum weight	____ pounds (range = 1 – 999)
	Maximum weight	____ pounds (range = 1 – 999)
	Length	__ feet (open-ended number)
	Width	__ feet (open-ended number)
	Height	__ feet (open-ended number)
	Percentage of time someone else helps push or pull a wheeled object	_____ % (range = 0 – 100)
	If percentage above is greater than zero: How many other people assist?	___ (range = 1 – 999)
	Surface	1 = hard surface (cement, macadam, tile, wood, or metal); 2 = dirt; 3 = gravel; 4 = padded floor; 5 = sand; 6 = other
	Duration of pushing or pulling	____ minutes (open-ended number; range = 1 – 999)
	Distance pushed or pulled	____ feet (open-ended number; range = 1 – 9999)

Ergonomic Category	Detailed Item/*Instructions*	Response Option
	Height pushed or pulled from	1 = above shoulder level; 2 = shoulder level; 3 = chest level; 4 = waist level; 5 = knee level; 6 = ankle or ground level; 7 = other
Pushing/pulling non-wheeled object		
	What is the non-wheeled object that is being pushed or pulled?	Open-ended text (50 characters maximum)
	Minimum weight	____ pounds (range = 1 – 999)
	Maximum weight	____ pounds (range = 1 – 999)
	Length	__ feet (open-ended number)
	Width	__ feet (open-ended number)
	Height	__ feet (open-ended number)
	Percentage of time someone else helps push or pull a non-wheeled object	_____ % (range = 0 – 100)
	If percentage above is greater than zero: How many other people assist?	___ (range = 1 – 999)
	Surface	1 = hard surface (cement, macadam, tile, wood, or metal); 2 = dirt; 3 = gravel; 4 = padded floor; 5 = sand; 6 = other
	Duration of pushing or pulling	____ minutes (open-ended number; range = 1 – 999)
	Distance pushed or pulled	____ feet (open-ended number; range = 1 – 9999)
	Height pushed or pulled from	1 = above shoulder level; 2 = shoulder level; 3 = chest level; 4 = waist level; 5 = knee level; 6 = ankle or ground level; 7 = other
Pushing/pulling on equipment parts		
	What is the equipment part that is being pushed or pulled?	Open-ended text (50 characters maximum)
	Minimum weight	____ pounds (range = 1 – 999)
	Maximum weight	____ pounds (range = 1 – 999)
	Length	__ feet (open-ended number)
	Width	__ feet (open-ended number)
	Height	__ feet (open-ended number)
	Percentage of time someone else helps push or pull an equipment part	_____ % (range = 0 – 100)

Ergonomic Category	Detailed Item/*Instructions*	Response Option
	If percentage above is greater than zero: How many other people assist?	___ (range = 1 – 999)
	Surface	1 = hard surface (cement, macadam, tile, wood, or metal); 2 = dirt; 3 = gravel; 4 = padded floor; 5 = sand; 6 = other
	Duration of pushing or pulling	____ minutes (open-ended number; range = 1 – 999)
	Distance pushed or pulled	____ feet (open-ended number; range = 1 – 9999)
	Height pushed or pulled from	1 = above shoulder level; 2 = shoulder level; 3 = chest level; 4 = waist level; 5 = knee level; 6 = ankle or ground level; 7 = other
Climbing		
	What is being climbed?	1 = ladder or scaffolding; 2 = hill or gulley; 3 = machinery; 4 = other
	Maximum vertical height climbed	____ feet (open-ended number; range = 1 – 999)
	If hill or gulley is being climbed: Surface type	1 = hard packed; 2 = sand; 3 = rock or scree; 4 = other
	If hill or gulley is being climbed: Length of hill or gulley	____ feet (open-ended number; range = 1 – 9999)
	Is equipment carried while climbing?	0 = no; 1 = yes
	If "Yes" above: Equipment weight when carrying and climbing	____ pounds (open-ended number; range = 1 – 9999)
Standing		
	Duration of continuously keeping this position	____ minutes (open-ended number; range = 1 – 999)
Non-standing		
	What is the most physically demanding non-standing position that must be used?	1 = stoop or squat (flex at knees and hips); 2 = lying down (on back, side, or stomach); 3 = kneel (on one or both knees on ground or surface); 4 = other
	Duration of continuously keeping this position	____ minutes (open-ended number; range = 1 – 999)
Walking		
	Maximum distance walked when completing this task	___ (open-ended number; range = 1 – 99999); _____ unit [1 = feet, 2 = miles, 3 = yards]
	Total time walking continuously	____ minutes (open-ended number; range = 1 – 999)

Ergonomic Category	Detailed Item/*Instructions*	Response Option
	Surface type	1 = hard surface (cement, macadam, tile, wood, or metal); 2 = dirt; 3 = gravel; 4 = padded floor; 5 = sand; 6 = other
	Surface slope degree or grade	0 = no slope; 1 = 9.5 degrees or 16.7% grade; 2 = 18.3 degrees or 33% grade; 3 = 33.7 degrees or 66% grade; 4 = other
	Walk up or down stairs	1 = No; 2 = Ascend only; 3 = Descend only; 4 = Ascend and Descend
	If Ascend and Descend selected above: How many <u>total</u> stairs must be walked up or down?	1 = 1 to 5 stairs; 2 = 6 to 10 stairs; 3 = 11 to 18 stairs or one floor; 4 = 19 to 35 stairs or two floors; 5 = 36 or more stairs or 3 or more floors
Running		
	Fastest running pace	1 = Jog; 2 = Run; 3 = Sprint; 4 = other
	Running distance	___ (open-ended number; range = 1 – 99999); _____ unit [1 = feet, 2 = miles, 3 = yards]
	Duration of run	____ minutes (open-ended number; range = 1 – 999)
	Surface type	1 = hard surface (cement, macadam, tile, wood, or metal); 2 = dirt; 3 = gravel; 4 = padded floor; 5 = sand; 6 = other
	Surface slope degree or grade	0 = no slope; 1 = 9.5 degrees or 16.7% grade; 2 = 18.3 degrees or 33% grade; 3 = 33.7 degrees or 66% grade; 4 = other
Crawling		
	Longest distance that must be crawled	____ feet (open-ended number; range = 1 – 999)
	Duration of crawl	____ minutes (open-ended number; range = 1 – 999)
Holding		
	In this task, what object is held that weighs at least 30 pounds?	Open-ended text (50 characters maximum; 1 response option)
	Minimum weight	____ pounds (range = 1 – 999)
	Maximum weight	____ pounds (range = 1 – 999)
	Length	__ feet (open-ended number)
	Width	__ feet (open-ended number)
	Height	__ feet (open-ended number)
	Percentage of time someone else helps with holding	_____ % (range = 0 – 100)
	If percentage above is greater than zero: How many other people assist?	___ (range = 1 – 999)

Ergonomic Category	Detailed Item/*Instructions*	Response Option
	Percentage of time held with two hands	_____ % (range = 0 – 100)
	Duration of hold	____ minutes (open-ended number; range = 1 – 999)
	Maximum height held at	1 = above shoulder level; 2 = shoulder level; 3 = chest level; 4 = waist level; 5 = knee level; 6 = ankle or ground level; 7 = other
Shoveling		
	Longest duration of shoveling required	____ minutes (open-ended number; range = 1 – 999)
	Material shoveled	1 = Dirt; 2 = Machinery or metal parts; 3 = Rock or gravel; 4 = Sand; 5 = Snow; 6 = other
Digging		
	Longest duration that digging is required	____ minutes (open-ended number; range = 1 – 999)
	Material dug up	1 = Earth (dirt, clay, or soil); 2 = Rocky soil; 3 = other
Pounding		
	Duration of pounding	____ minutes (open-ended number; range = 1 – 999)
	Pounding tool	1 = Hammer; 2 = Mallet; 3 = Sledgehammer; 4 = Slugging wrench; 5 = Tamper; 6 = other
Using powered hand-held tools		
	Duration of continuous use of a powered handheld tool	____ minutes (open-ended number; range = 1 – 999)
	Tool weight	____ pounds (open-ended number; range = 1 – 999)
Using non-powered hand-held tools		
	Duration of continuous use of a non-powered hand-held tool	____ minutes (open-ended number; range = 1 – 999)
	Tool weight	1 = 1 to 3 pounds; 2 = 4 to 6 pounds; 3 = 7 to 10 pounds

Appendix J: Additional Information About HumRRO's Criterion-Related Validation Study Efforts

This appendix provides additional background and technical information provided to RAND by HumRRO. These additional details are provided to supplement and provide context for the analyses RAND conducted and presented in Chapter Six.

An Overview of the Criterion-Related Validation Study

Using job analysis data from interviews and site visits as described in Chapter Four, HumRRO designed and executed a criterion-related validation study to identify fitness tests that can determine an individual's capability to perform physically demanding tasks required by a range of AFSCs. The following sections summarize the (a) development of the physical task simulations to measure physical performance on job-relevant tasks, (b) physical tests used in the validation study, (c) statistical relationships between physical tests and task simulations, and (d) explored efforts to establish SAT minimum scores for each career field (ultimately unsuccessful due to limitations of the available data).

How Can Physical Performance on Job-Relevant Tasks Be Measured?

Before the minimum requirements on a test can be determined, performance measures must be developed that can be used to identify the test score(s) associated with minimally acceptable performance. This connection is generally established using data collected from a criterion-related validation study in which individual test scores are correlated with individual job or task performance. Physical performance can be measured using a variety of methods including subjective rating measures, such as supervisor ratings of performance and peer ratings of performance, and more objective measures, such as task simulations. Existing measures of job performance (e.g., EPR) do not sufficiently measure an individual's physical performance. Although subjective rating measures could have been created for the purpose of this study, task simulations are a more direct, objective measure of physical performance because they approximate the physical demands of job-relevant tasks (e.g., Henderson, 2010; Williford et al., 1999).

Development of Physical Task Simulations

The first step involved reviewing the job analysis results from the MCQ to determine movement categories required by most of the 21 specialties included in the job analysis. These movement categories help to identify common physical requirements among AFSCs performing

very different tasks. For example, lifting and carrying may be common physical demands for tasks performed by "Cable and Antenna Systems" during maintenance and installation of antennas and cables. These movement categories may also be required to remove or install bomb racks in "Aircraft Armament Systems." Although the tasks are quite different, both require muscular strength to meet the lifting and carrying demands associated with both tasks.

This review found that many of the ten physically demanding tasks within an AFS contained the same movement categories across the 21 AFSCs. Furthermore, the movement categories of Lift, Carry, Push/Pull, Climb, Stand, Nonstand/Kneel, Hold, and Operate Nonpower Tools are linked to the 75 percent or more of the AFSs (19 to 21), while Walk, Crawl, Pound, and Operate Power Tools are linked to approximately 50 percent of the AFSCs. The remaining movement categories (Run, Shovel, Dig, Swim) were linked to less than 50 percent of the AFSCs.

Movement categories linked to the 75 percent or more of the AFSCs were retained for physical task simulation development. The results yielded eight potential task simulations. Following further review of task simulations and to stay within the four-hour time constraint for testing, the number of task simulations was reduced to four. These four task simulations were selected to be the most representative of the physical demands required by AFSCs sampled for the study:

- **Lift and Carry**: Lift equipment associated with multiple AFSCs from six inches (ankle/ground level) to 72 inches (above shoulder level). The objects were moved from one platform to another across a 15- to 30-foot distance. After completing the movement of all objects, the airman repeated the process, and all objects were moved back to their starting positions. This completed one cycle of the Lift/Carry physical task simulation. A total of two cycles was completed in this physical task simulation.
- **Push and Pull**: Push portable lights, then push a tool chest and a portable heater. After each of the three objects was pushed, each object was then pulled back to its original position. This completed one cycle. A total of two cycles was completed.
- **Climb and Carry**: Climb an extension ladder to a height of 9 feet while wearing a 30-pound vest. When an airman's feet reached the ninth rung (9 feet), the airman moved a small simulated tool box (14 pounds) from one location to another. The movement of the box simulated the airmen performing a task similar to moving equipment and parts during installation and removal tasks. The Climb physical task simulation began by carrying the ladder 45 feet, similar to removing it from storage and carrying it to the work location. The ladder was placed on the ground and the airman ascended an extension ladder affixed to a pillar. Upon reaching the 9-foot level, the airman moved the simulated tool box from one location to another and descended the ladder. This completed one cycle of the Climb. Four cycles were completed for this physical task simulation.
- **Hold:** Three cycles of holding objects of varying weights at different levels. The first cycle involved holding five objects of varying weight at chest level while in a standing position. The second cycle involved holding the same five objects above shoulder level for 20 seconds, followed by a ten-second rest period. The third cycle used the same protocol but was performed from a squatting position.

Table J.1 summarizes the relevance of each physical task simulation to each AFSC included in the job analysis. The table shows that the four task simulations were relevant to most AFSCs. (Nonrelevant task simulations are noted by a boldface "no" in the table.)

In addition to pilot testing each of the task simulations, we collected data from airmen to determine the test-retest reliability of the task simulations, which helps to increase confidence that physical performance is relatively stable and not easily learned nor influenced by extraneous factors such as equipment malfunctions. Test-retest reliability coefficients are computed by administering a test to a group of participants at two times and then correlating their two sets of scores. The obtained correlation coefficient indicates how similar the scores of the same group of participants are over the two administrations of the same measure. Test-retest reliability coefficients range from –1.00 to 1.00, with 1.00 indicating perfect reliability. The test-retest correlations for the four task simulations were high and ranged from 0.77 (Hold) to 0.93 (Lift/Carry). The high correlations indicate that these four task simulations are consistent measures of physical task performance (Dancey and Reidy, 2004).

Table J.1. Linkage of Task Simulations to AFSCs

			Task Simulations by Movement Category			
Job#	**AFSC**	**Specialty Title**	**Lift/Carry**	**Push/Pull**	**Climb**	**Hold**
1	1A0X1	In-Flight Refueling	Yes	Yes	Yes	Yes
2	1A2X1	Aircraft Loadmaster	Yes	Yes	Yes	Yes
3	2A3X3L	Tactical Aircraft Maintenance	Yes	Yes	Yes	Yes
4	2A5X2	Helicopter/Tiltrotor Aircraft Maintenance	Yes	Yes	Yes	Yes
5	2A6X1	Aerospace Propulsion	Yes	Yes	Yes	Yes
6	2A6X2	Aerospace Ground Equipment	Yes	Yes	Yes	No
7	2A6X3	Aircrew Egress Systems	Yes	Yes	Yes	Yes
8	2A7X1	Aircraft Metals Technology	Yes	Yes	Yes	Yes
9	2F0X1	Fuels	Yes	**No**	Yes	**No**
10	2M0X2	Missile and Space Systems Maintenance	Yes	Yes	Yes	Yes
11	2S0X1	Materiel Management	Yes	Yes	Yes	**No**
12	2W0X1	Munitions Systems	Yes	Yes	Yes	Yes
13	2W1X1E	Aircraft Armament Systems	Yes	Yes	Yes	Yes
14	3D1X7	Cable and Antenna Systems	Yes	Yes	Yes	Yes
15	3E1X1	Heating, Ventilation, Air Conditioning, and Refrigeration	Yes	Yes	Yes	Yes
16	3E2X1	Pavements and Construction Equipment	Yes	Yes	Yes	Yes
17	3E4X1	Water and Fuel Systems Maintenance	Yes	Yes	Yes	Yes
18	3E7X1	Fire Protection	Yes	Yes	Yes	**No**
19	3E8X1	Explosive Ordnance Disposal	Yes	Yes	Yes	Yes
20	3P0X1	Security Forces	Yes	No	Yes	**No**
21	4B0X1	Bioenvironmental Engineering	Yes	Yes	Yes	**No**

.

What Physical Fitness Tests Were Considered?

Physical fitness can be measured in a number of ways using tests of various abilities (McArdle, Katch, and Katch, 2010). To select tests for this study, past validation research was reviewed to identify physical tests that assessed the relevant abilities significantly related to measures of job performance (e.g., Blakley et al., 1994; Gebhardt and Baker, 2010a; Rayson, Holliman, and Belyavin, 2000). This review identified a variety of physical tests. It also showed

that tests of flexibility (e.g., Sit and Reach, Twist and Touch) were rarely related to job performance (e.g., Gebhardt and Baker, 2010b). Thus, measures of flexibility were not included in the test development. Additionally, manual dexterity was eliminated from further consideration because it was identified for only two of the 16 movement categories.

Following this review, 19 tests were identified as possible candidates for the validation study. However, several tests were further discarded because of constraints in the time allotted to test research participants. Similarly, other tests were eliminated because of potential space or time restrictions at the MEPSs, where future testing would take place. Descriptions and abilities measured by each of the final nine fitness tests are provided below and in Table J.2.

Arm Endurance

The Arm Endurance test measures the ability of the muscles of the upper body to exert force repeatedly or continuously over a moderate time period. Thus, this test measures anaerobic power and muscular endurance. The Arm Endurance test involves pedaling a stationary arm ergometer with the arms for one minute with a fixed workload (i.e., resistance). The test is scored as the number of revolutions pedaled in one minute. This test has been found to be a valid predictor of job performance in a number of validation projects with validity coefficients ranging from 0.21 to 0.72 (Gebhardt and Baker, 2010b).

Arm Lift

The Arm Lift test measures strength in the upper body. It evaluates the maximum force that one can exert for a brief time period. The test involves generating a steady maximal force in an upward direction with the elbows flexed at 90 degrees. Three trials are given. This test has been used for the selection of workers for public safety, materials handling, and maintenance jobs. It has been found to be statistically reliable and a valid predictor of job performance with validity coefficients as high as 0.74 (Chaffin et al., 1977; Gebhardt and Baker, 2010b).

Handgrip

The Handgrip test measures grip strength. Three trials are given for both the dominant and nondominant hands. A mean is calculated for the dominant and nondominant hand trials, along with a mean of the six trials. The Handgrip test was found to be a valid predictor of job performance when the criterion measure included activities such as lifting and pulling (e.g., $r = 0.63$) (Gebhardt and Baker, 2010b).

Table J.2. Abilities Measured by Each Test

	Physical Ability					
Test	Muscular Strength	Muscular Endurance	Aerobic Capacity	Equilibrium	Anaerobic Power	Coordination
Arm Endurance	No	Yes	No	No	No	No
Arm Lift	Yes	No	No	No	No	No
Handgrip	Yes	No	No	No	No	No
Plank Test	Yes	Yes	No	No	No	No
Push-Ups	Yes	Yes	No	No	No	No
SAT (Strength Aptitude Test)	Yes	Yes	No	No	No	No
Sit-Ups	No	Yes	No	No	No	No
Standing Broad Jump	Yes	No	No	No	Yes	Yes
Step Test	No	Yes	Yes	No	No	No

Plank Test

The Plank test assesses trunk strength. The test position is with the toes and forearms on a mat and the legs, buttocks, and back in straight alignment. The position is held for as long as possible. When the chest touches the mat or the legs, buttocks, or back are not in straight alignment, the test is completed. The score is the time the position is held. The Plank test is a valid and reliable measure of global core muscular endurance (Baker and Gebhardt, 2012).

Push-Ups

Push-ups measure upper body muscular strength and muscular endurance. The test involves performing as many push-ups as possible in one minute, using correct form. The test is started in the extended, or up, position. A completed push-up is defined as lowering the body to the point at which the sternum/chest touches a foam block and returning to the start position. The score is the number of push-ups completed in one minute. Push-ups have been found to be a valid predictor of job performance for law enforcement positions with validity coefficients ranging from 0.34 to 0.81 (Gebhardt and Baker, 2010b).

Sit-Ups

Sit-ups measure muscular endurance of the abdominal musculature. This test is performed with the knees flexed and the arms held across the chest. The score is the number of sit-ups completed in one minute. Sit-ups have been found to be a valid predictor of job performance for public safety jobs and jobs in the railroad, freight, natural gas, and telecommunication industries with validity coefficients up to 0.68 (Gebhardt and Baker, 2010b).

Standing Broad Jump

This test is used to measure primarily anaerobic power. The test begins with the individual standing behind a line marked on the ground with feet slightly apart. A two-foot takeoff and landing is used, with swinging of the arms and flexing of the knees to provide forward propulsion (Koch et al., 2003). The goal is to jump as far as possible, landing on both feet without falling backwards. The distance of the jump will be measured from the starting line to the back of the feet. Three trials will be given. The Standing Broad Jump was a valid predictor of firefighter performance (Dotson et al., 1978), as well as highly related (r = 0.81–0.84) to carrying objects and stretchers (Bilzon et al., 2003).

Step Test

This test is used to measure aerobic capacity. This test involves stepping up and down on a platform at a specified cadence (96 steps per minute) for a total of three minutes. The participant's heart rate is taken following the completion of the test to determine the individual's aerobic fitness. Step tests have been found to be valid predictors of job performance for manual materials handling and law enforcement positions (Gebhardt and Baker, 2010b).

Strength Aptitude Test (SAT)

Revisions to the SAT protocol were made for the validation study. The first revision was to have participants continue with the test after they successfully lifted 110 pounds. If the lift of 110 pounds was successful, 10 pounds was added and the test continued. Participants continued the incremental lifts until either (a) they could not make a successful lift or (b) they successfully lifted 190 pounds. One hundred and ninety pounds was selected as the SAT's endpoint because the ILM's maximum was 190 pounds. This change was made to increase the variability of scores and obtain a more accurate measure of an individual's muscular strength.

The second protocol change involved one additional lift after the participant's unsuccessful lift. The revision to the protocol specified that after an unsuccessful lift, 5 pounds was removed and a final lift was attempted. This was the participant's final lift regardless of whether the lift was successful or not. If this final lift was successful, the participant's score was the weight of this final lift. If this final lift was not successful, the participant's score was the weight of the final successful lift. This change was made in an attempt to obtain a more precise estimate of physical strength on the SAT.

Which Physical Fitness Tests Are Valid Indicators of an Individual's Capability to Meet Job-Relevant Physical Demands?

To evaluate the predictive validity of the fitness tests, physical task simulation scores were first combined into a single composite score to provide an overall measure of physical job

performance. That is, the composite physical task simulation score was the sum of standardized scores on the four task simulations: (a) Climb and Carry, (b) Hold, (c) Lift and Carry, and (d) Push and Pull. After composite physical task simulation scores were computed, scores on each fitness test were correlated with scores on the physical task simulation composite.

Validity of Individual Physical Fitness Tests

All of the fitness tests in the study were significantly correlated ($p < 0.05$) with the physical task simulation composite (see Table J.3). In addition to its significant correlation with the composite, the SAT had significant correlations with all four individual task simulations. These significant correlations demonstrate that the SAT predicts physical job performance, which indicates that higher SAT scores are associated with better physical task simulation performance. Additional statistical analyses on these tests and possible test combinations were conducted by RAND and are presented in Chapter Six.

Table J.3. Correlations Between Fitness Tests and Task Simulations

Fitness Test	Climb and Carry Final Time	Hold All Cycles Total Time Held	Lift/Carry Final Time	Push/Pull Final Time	Composite (All Task Simulations)
Arm Endurance	–0.61	0.66	–0.69	–0.69	0.80
Arm Lift Mean (3 Trials)	–0.49	0.73	–0.59	–0.60	0.72
Handgrip Total (3 Trials)	–0.56	0.71	–0.65	–0.67	0.77
Plank Test	–0.19	0.29	–0.20	–0.17	0.27
Push-Ups	–0.41	0.68	–0.47	–0.53	0.62
Sit-Ups	–0.30	0.43	–0.34	–0.37	0.42
Standing Broad (3 Trials)	–0.48	0.62	–0.58	–0.59	0.67
Step Test VO_2	–0.20	0.27	–0.28	–0.21	0.29
Strength Aptitude Test	–0.51	0.76	–0.60	–0.68	0.76

NOTE: All correlations are significant at $p < 0.01$.

Establishing Minimum SAT Requirements

HumRRO followed several steps in exploring whether an updated formula for determining the minimum SAT requirements for each AFSC could be developed. The first step was to clean the job analysis data from surveys, interviews, and site visits. This step required the identification of incomplete responses, elimination of extreme responses (e.g., lifting over 500 pounds), and elimination of non–job-related tasks and equipment (e.g., lifting fitness equipment when AFSC was clearly not fitness-related). Once HumRRO had cleaned the data, they aggregated responses by computing the average across responses within each AFSC (e.g., mean of maximum weight lifted). The objective was to cluster AFSCs using different combinations of job analysis variables; however, the cluster analyses explored by HumRRO produced varying

results, depending on which job analysis variables were included. Such varying results can be expected from cluster analysis, which can find similarities among AFSCs that do not necessarily affect the strength demands. For example, one of the cluster analysis models grouped AFSCs based on the presence or absence of ladders, but not all AFSCs that require climbing ladders have similar strength demands.

After reviewing HumRRO's clustering efforts, it became clear to RAND that to increase the accuracy of grouping AFSCs by their physical demand, there needs to be additional criteria for determining how well a particular clustering solution works. For example, one way to establish such criteria is to use SMEs who are familiar with all of the AFSCs in the Air Force and can rank or group AFSCs by physical demand. Given the number and range of AFSCs in the Air Force, this strategy was not feasible. In the future, the Air Force should gather physical performance data from personnel assigned to physically demanding AFSCs to determine if the current SAT requirement is sufficient or if it needs to be changed. Furthermore, trained exercise science analysts and senior leaders familiar with the physical demands of multiple AFSCs can review the recommended SAT requirements to ensure that the more–physically demanding AFSCs have higher SAT requirements compared with less–physically demanding AFSCs. RAND's discussion of other strategies for establishing these clusters is presented in Chapter Four.

Appendix K: Technical Background for Additional RAND Analyses

With eight possible predictors in addition to the SAT (which is to be used in every model), there are 256 possible subsets of predictors to consider. Consequentially, it is feasible to fit each possible model and compare their respective results. Models are compared using Akaike's information criterion (AIC). For a specific model, this criterion is calculated via

$$\text{AIC} = n\log(SSE) - n\log(n) + 2p,$$

where $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the sum of squared errors, with $\hat{y}_i$ denoting the predicted value of y_i, the outcome for individual i. It is commonly understood that superior models are ones that yield smaller values of *SSE*; however, inclusion of unnecessary predictors can spuriously deflate *SSE*. Therefore, the term $2p$ within the formula for AIC is used to penalize models that have a larger number of predictors. The best model is the one with the lowest value of *AIC*. An alternative criterion that can be used to find the adequacy of any fitted model is the predicted residual sum of squares (PRESS) statistic, which is a form of cross-validation. For a specific model, this statistic is calculated using the following formula:

$$\text{PRESS} = \sum_{i=1}^{n}(y_i - \hat{y}_{i,-i})^2,$$

where $\hat{y}_{i,-i}$ is the predicted value of the outcome for observation i when the model is calculated while excluding data for the i^{th} observation (and including all other observations). A model selection strategy that fits models using all possible subsets of predictors and then selects an optimal model on the basis of either AIC or the PRESS statistic is preferable to stepwise selection (due to its exhaustive consideration of possible combinations of predictors and due to the rigorous evaluation the AIC and PRESS criteria have received in the scientific literature).

One of our objectives in selecting tests is to minimize differential validity (i.e., the presence of predictive bias) on the basis of gender. That is, it is possible that the predicted outcomes for men will systematically underestimate the true outcome value, whereas the predicted outcomes for women will systematically overestimate the true outcome value—we want to remove such biases from our predictions. We assess differential validity on the basis of gender as follows. For a given model, let $\mathbf{x}_i$ denote the vector of predictors for ($\mathbf{x}_i$ may be of dimension 1 to 9) and let y_i denote the outcome for individual i. Further, let s_i be a binary variable indicating whether or not individual i is male. The various test batteries are compared by assessing the validity of models of the form

$$\text{Regression (a): } y_i = \beta_0 + \boldsymbol{\beta}_1'\mathbf{x}_i + \epsilon_i,$$

where the β terms are regression coefficients, and ϵ_i is a mean-zero error term. To assess whether or not including gender in the above model will improve validity, we consider the following:

$$\text{Regression (b): } y_i = \beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i + \beta_2 s_i + \boldsymbol{\beta}_3' (s_i \star \mathbf{x}_i) + \epsilon_i,$$
$$\text{Regression (c): } y_i = \beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i + \boldsymbol{\beta}_3' (s_i \star \mathbf{x}_i) + \epsilon_i,$$
$$\text{Regression (d): } y_i = \beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i + \beta_2 s_i + \epsilon_i,$$

where $s_i \star \mathbf{x}_i$ represents the (multivariate) interaction of gender with the test battery. Predictive bias is gauged by considering the statistical significance of various subsets of coefficients in Regression (b). Specifically, to assess the presence of any gender effects, we compare Regression (b) to Regression (a) by testing $H_0: \beta_2 = 0$ and $\boldsymbol{\beta}_3 = \mathbf{0}$ against $H_1: \beta_2 \neq 0$ or $\boldsymbol{\beta}_3 \neq \mathbf{0}$. To test for slope effects, we compare Regression (b) to (d) by testing $H_0: \boldsymbol{\beta}_3 = \mathbf{0}$ against $H_1: \boldsymbol{\beta}_3 \neq \mathbf{0}$. The form of the test used to assess intercept effects is dependent upon the presence of slope effects. Specifically, if there are slope effects (i.e., if p-value ≤ 0.05 for the slope effects test), we test for intercept effects by comparing Regression (b) to (c) with $H_0: \beta_2 = 0$ and $H_1: \beta_2 \neq 0$. If there are not slope effects, we employ a test that incorporates that information: We test for intercept effects by comparing Regression (d) to (a) with $H_0: \beta_2 = 0$ and $H_1: \beta_2 \neq 0$. Each comparison is performed by applying an omnibus hypothesis test to the requisite set of estimated regression coefficients as found using Regression (b). This scheme for assessing differential validity is motivated by the exposition of Lautenschlager and Mendoza (1986). We report *p*-values for each of these hypothesis tests.

We also wish to account for the cost of implementation of the various tests. Instead of developing an objective function for optimization that in some arbitrary manner balances model validity, predictive bias, and cost, we focus our selection of the optimal model on the basis of model validity as measured by AIC. However, to ensure that the variety of objectives is addressed, we identify a range of optimal models. Note that, to simplify the discussion, the cost of a test is appraised only on the basis of whether the test is expensive or inexpensive. Continuing, we evaluated the combination of tests that could address the following questions: Which single test and, likewise, which combination of tests have the highest incremental validity beyond the SAT alone? Similarly, which low-cost test and which combination of low-cost tests have the highest incremental validity beyond the SAT alone? For the sake of comparison, we also consider an option that includes SAT as the sole predictor. In that vein, our discussion is reduced to comparison of five separate test batteries, outlined as follows:

- Option 1: SAT is the only predictor used (baseline).
- Option 2: SAT plus any single predictor
- Option 3: SAT plus as many other predictors as needed
- Option 4: SAT plus any single inexpensive test
- Option 5: SAT plus all inexpensive tests.

The predictors for Options 2, 3, and 4 are selected as the best-fitting set of predictors that satisfy the specific criteria. The following steps are taken to determine best-fitting test batteries: Each of the five outcome variables (which include four distinct outcomes and a composite outcome, calculated as a standardized average of the other four) is modeled using each test

battery. Specifically, we fit Regression (a) as defined above; i.e., gender is excluded from models used in the selection of the best-fitting test battery. The AIC value (along with the PRESS statistic) is calculated each time a model is fitted. Note that we must assess the validity of each predictor set across four separate outcomes. Hence, the best set of predictors is the one that has the lowest of the resulting AIC values for the respective model of the composite outcome. Models are also selected in a similar manner on the basis of the PRESS statistic for comparison. In addition to AIC and PRESS, we also store the R-squared value for each fitted model. Although the models are selected on the basis of predictive validity and costs, we also evaluate the extent to which test bias with respect to gender exists for each of these different test combinations.

Results are shown in Table K.1, wherein models have been selected using AIC. First, the table indicates whether or not each predictor is included in each of the various models. Next, the table gives the value of R-squared for each specific outcome including the composite (and when averaged across the four distinct outcomes). Then, for each outcome, the table gives the percentage change in R-squared (from the R-squared given when only SAT is included in the model) that is yielded by the fit of the respective test battery. Next, the table gives the *p*-value for the test of gender differences (i.e., a comparison of Regression [b] to Regression [a]), the *p*-value for the test of gender-based intercept differences (i.e., a comparison of Regression [b] to Regression [c]), and the *p*-value for the test of gender-based slope differences (i.e., a comparison of Regression [b] to Regression [d]). For each type of test (the tests for overall differences, the tests for intercept differences and the tests for slope differences), we also provide *p*-values derived using an omnibus test (e.g., we simultaneously test for gender differences across the four distinct outcomes).

Table K.1. Results from the Comparison of Test Batteries

		Option 1	Option 2	Option 3	Option 4	Option 5
Physical Fitness Test	SAT	X	X	X	X	X
	Arm Endurance	—	X	X	—	—
	Push-Ups	—	—	X	—	X
	Sit-Ups	—	—	—	—	X
	Arm Lift*	—	—	X**	—	—
	Handgrip*	—	—	X	—	—
	Plank Test	—	—	—	—	X
	Standing Broad Jump*	—	—	—	X	X
	Step Test	—	—	—	—	—
R-squared	Climb Task Simulation	0.262	0.383	0.413	0.280	0.283
	Hold Task Simulation	0.572	0.597	0.651	0.578	0.615
	Lift and Carry Task Simulation	0.328	0.447	0.483	0.363	0.367
	Push and Pull Task Simulation	0.463	0.545	0.573	0.477	0.479
	Standardized Composite (all task simulations)	0.584	0.703	0.746	0.609	0.617
	Average (four task simulations)	0.406	0.493	0.530	0.424	0.436
Percentage change in R-squared from SAT only	Climb Task Simulation	—	46.45	57.91	7.04	8.38
	Hold Task Simulation	—	4.26	13.67	1.00	7.44
	Lift and Carry Task Simulation	—	36.16	47.18	10.57	11.85
	Push and Pull Task Simulation	—	17.60	23.60	2.94	3.38
	Standardized Composite (all task simulations)	—	20.27	27.65	4.18	5.59
	Average (four task simulations)	—	26.12	35.59	5.39	7.72
p-value for Gender effects	Climb Task Simulation	6.18E-04	0.669	0.026	0.027	0.02
	Hold Task Simulation	0.002	0.327	0.839	0.027	0.291
	Lift and Carry Task Simulation	2.49E-10	0.005	0.022	1.35E-06	8.38E-07
	Push and Pull Task Simulation	5.84E-13	3.48E-05	1.21E-03	2.56E-10	3.78E-09
	Standardized Composite (all task simulations)	8.13E-14	1.73E-03	0.009	1.46E-09	3.98E-08
	Omnibus	0.000	1.17E-04	6.88E-04	2.53E-10	1.04E-09

		Option 1	Option 2	Option 3	Option 4	Option 5
p-value for slope effects	Climb Task Simulation	0.024	0.463	0.019	0.106	0.065
	Hold Task Simulation	0.868	0.838	0.742	0.618	0.806
	Lift and Carry Task Simulation	7.14E-04	0.075	0.048	0.007	0.003
	Push and Pull Task Simulation	1.64E-05	0.004	0.016	8.61E-05	5.73E-04
	Standardized Composite (all task simulations)	1.14E-04	0.05	0.011	7.63E-04	0.004
	Omnibus	4.08E-05	0.034	0.013	1.37E-03	5.80E-04
p-value for intercept effects	Climb Task Simulation	2.00E-03	0.861	0.236	0.012	0.033
	Hold Task Simulation	0.19	0.823	0.702	0.167	0.078
	Lift and Carry Task Simulation	3.21577E-07	0.103	0.023	3.38E-04	1.90E-03
	Push and Pull Task Simulation	8.68108E-10	0.002	6.01E-04	1.05227E-06	5.42139E-07
	Standardized Composite (all task simulations)	5.04841E-09	0.05	0.006	3.51001E-06	8.98985E-06
	Omnibus	9.04565E-11	0.02	0.003	4.06757E-06	6.07139E-06

* Test scores were computed using the mean of three trials.
** If the PRESS statistic is used as the criterion for model selection, Option 3 does not include Arm Lift—the options are otherwise the same, however, when the PRESS criterion is used.

The results indicate that Arm Endurance adds the most validity of any predictor (see Option 2). Comparison of Option 1 to Option 2 indicates that adding Arm Endurance to SAT gains a fairly substantial improvement. Specifically, we see a 20.3-percent increase in R-squared for the composite outcome. Furthermore, when comparing Option 2 with Option 3, we see that improvement in validity can be gained by adding other predictors beyond Arm Endurance; however, this improvement is not as substantial (i.e., the R-squared increases only from 0.703 to 0.746 for the composite outcome). When examining Options 4 and 5, we see that inclusion of inexpensive tests offers only minor improvements in validity. Specifically, including all four inexpensive tests in addition to SAT increases R-squared only from 0.584 to 0.617 for the composite outcome.

We see that there is strong evidence (in all outcomes) that there is statistically significant predictive bias because of gender differences (in all outcomes) when only SAT is used as a predictor (i.e., the *p*-value for gender effects in Option 1 is $8.14 \cdot 10^{-14}$ for the composite outcome). However, when Arm Endurance and other predictors are used in addition to SAT, we see that a good portion of this predictive bias is alleviated, although it is not removed entirely (this is evident graphically in Figures 6.1 and 6.2). We see that inclusion of inexpensive tests is mostly ineffective at reducing predictive gender biases (i.e., the *p*-values for gender effects for

the composite outcome with Options 4 and 5 are $1.46 \cdot 10^{-9}$ and $3.98 \cdot 10^{-8}$, respectively). Furthermore, when only the SAT is included in the models, we see stronger evidence of intercept effects than slope effects. When we expand the test battery to include tests with additional equipment costs, we no longer see stronger evidence of intercept effects than slope effects (or vice-versa, for that matter). However, if the test battery is allowed to include only tests with no equipment costs in addition to the SAT, we again see more evidence of intercept effects than of slope effects.

References

AFECD—*See Air Force Enlisted Classification Directory*.

Air Force Enlisted Classification Directory, Washington, D.C.: Headquarters, Air Force Personnel Classification, April 30, 2013.

Ayoub, M. M., B. C. Jiang, J. L. Smith, J. L. Selan, and J. W. McDaniel, "Establishing a Physical Criterion for Assigning Personnel to U.S. Air Force Jobs," *American Industrial Hygiene Association Journal*, Vol. 48, No. 5, May 1987, pp. 464–470.

Baker, T. A., and D. L. Gebhardt, "Chapter 13: The Assessment of Physical Capabilities in the Workplace," in N. Schmitt, ed., *Handbook of Assessment and Selection*, New York: Oxford University Press, 2012, pp. 274–296.

Bilzon, J. L. J., E. G. Scarpetto, E. Bilzon, and A. J. Alsopp, *Generic Task-Related Occupational Requirements for Royal Navy (RN) Personnel*, UK Ministry of Defense, 2003.

Blakley, Barry R., Miguel A. Quiñones, Marnie Swerdlin Crawford, and I. Ann Jago, "The Validity of Isometric Strength Tests," *Personnel Psychology*, Vol. 47, No. 2, June 1994, pp. 247–274.

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

Chaffin, D. B., G. D. Herrin, W. M. Keyserling, and A. Garg, "A Method for Evaluating the Biomechanical Stresses Resulting from Manual Materials Handling Jobs," *American Industrial Hygiene Association Journal*, Vol. 38, No. 12, December 1977, pp. 662–675.

Cizek, G. J. (ed.), *Setting Performance Standards: Foundations, Methods, and Innovations*, London: Routledge, 2012.

Copley, G. B., B. R. Burnham, M. J. Shim, and P. A. Kemp, "Using Safety Data to Describe Common Injury-Producing Events: Examples from the U.S. Air Force," *American Journal of Preventive Medicine*, Vol. 38, No. 1, 2010, pp. S117–S125.

Crocker, Linda, and James Algina, *Introduction to Classical and Modern Test Theory*, Belmont, Calif.: Wadsworth Publishing, 1986.

Dancey, Christine P., and John Reidy, *Statistics Without Maths for Psychology: Using SPSS for Windows*, Upper Saddle River, N.J.: Prentice Hall, 2004.

DoD—*See* U.S. Department of Defense.

Dotson, C. O., D. L. Santa Maria, P. O. Davis, and R. A. Schwartz, *The Development of Job-Related Physical Performance Examinations for Fire Fighters*, Washington, D.C.: U.S. Government Printing Office, Stock No. 003-000-00541-1, 1978.

GAO—*See* U.S. General Accounting Office.

Gebhardt, Deborah L., and Todd A. Baker, "Physical Performance Tests," in J. L. Farr and N. T. Tippins, eds., *Handbook of Employee Selection*, New York: Routledge/Taylor & Francis Group, 2010a, pp. 277–298.

———, "Physical Performance," in J. C. Scott and D. H. Reynolds, eds., Handbook of Workplace Assessment, San Francisco, Calif.: Jossey-Bass, 2010b, pp. 179–196.

Gebhardt, D. L., T. A. Baker, and A. Thune, *Development and Validation of Physical Performance, Cognitive, and Personality Assessments for Selectors and Delivery Drivers*, Beltsville, Md.: Human Performance Systems, 2006.

Henderson, N. D., "Predicting Long-Term Firefighter Performance from Cognitive and Physical Ability Measures," *Personnel Psychology*, Vol. 63, No. 4, 2010, pp. 999–1039.

Hoffman, C. C., "Generalizing Physical Ability Test Validity: A Case Study Using Test Transportability, Validity Generalization, and Construct Related Validation Evidence," *Personnel Psychology*, Vol. 52, No. 4, 1999, pp. 1019–1041.

Hoffman, C. C., L. M. Holden, and K. Gale, "So Many Jobs, So Little 'N': Applying Expanded Validation Models to Support Generalization of Cognitive Test Validity," *Personnel Psychology*, Vol. 53, No. 4, 2000, pp. 955–991.

Knapik, J. J., B. H. Jones, M. A. Sharp, S. Darakjy, S. Jones, K. G. Hauret, and G. Piskator, "The Case for Pre-Enlistment Physical Fitness Testing: Research and Recommendations," 12-HF-01Q9D-04, Aberdeen Proving Ground, Md.: U.S. Army Center for Health Promotion and Preventive Medicine, 2004.

Knapik, J. J., J. E. Wright, D. M. Kowal, and J. A. Vogel, "The Influence of U.S. Army Basic Initial Training on the Muscular Strength of Men and Women," No. USARIEM-M-11/80, Natick, Mass.: Army Research Institute of Environmental Medicine, 1980.

Koch, A. J., H. S. O'Bryant, M. E. Stone, K. Sanborn, C. Proulx, J. Hruby, E. Shannonhouse, R. Boros, and M. H. Stone, "Effect of Warm-Up on the Standing Broad Jump in Trained and Untrained Men and Women," *Journal of Strength Conditioning Research*, Vol. 17, No. 4, November 2003, pp. 710–714.

Lautenschlager, G. J., and J. L. Mendoza, "A Step-Down Hierarchical Multiple Regression Analysis for Examining Hypotheses About Test Bias in Prediction," *Applied Psychological Measurement*, Vol. 10, No. 2, 1986, pp. 133–139.

McArdle, William D., Frank I. Katch, and Victor L. Katch, *Exercise Physiology: Nutrition, Energy, and Human Performance*, Philadelphia, Pa.: Lippincott Williams & Wilkins, 2010.

McDaniel, J. W., R. J. Skandis, and S. W. Madole, *Weight Lift Capabilities of Air Force Basic Trainees*, AFAMRL Report TR-83-0001, Wright-Patterson Air Force Base, Ohio: Air Force Aerospace Medical Research Laboratory, 1983.

Messing, K., and J. Stevenson, "Women in Procrustean Beds: Strength Testing and the Workplace," *Gender, Work & Organization*, Vol. 3, No. 3, 1996, pp. 156–167.

Myers, D. C., D. L. Gebhardt, and C. E. Crump, *Validation of the Military Entrance Physical Strength Capacity Test*, Technical Report No. 610, Alexandria, Va.: U.S. Army Research Institute for the Behavioral and Social Sciences, 1984.

Public Law 113-291, Carl Levin and Howard P. "Buck" McKeon National Defense Authorization Act for Fiscal Year 2015, Section 524, December 19, 2014.

Rayson, M. P., D. E. Holliman, and A. Belyavin, "Development of Physical Selection Procedures for the British Army, Phase 2: The Relationship Between Physical Performance Tests and Criterion Tasks," *Ergonomics*, Vol. 43, 2000, pp. 73–105.

Roth, P. L., H. Le, I. S. Oh, C. H. Van Iddekinge, M. A. Buster, S. B. Robbins, and M. A. Campion, "Differential Validity for Cognitive Ability Tests in Employment and Educational Settings: Not Much More Than Range Restriction?" *Journal of Applied Psychology*, Vol. 99, No. 1, 2014, pp. 1–20.

Scherbaum, C. A., "Synthetic Validity: Past, Present, and Future," *Personnel Psychology*, Vol. 58, No. 2, 2005, pp. 481–515.

Sims, Carra S., Chaitra M. Hardison, Maria C. Lytell, Abby Robyn, Eunice C. Wong, and Erin Gerbec, *Strength Testing in the Air Force: Current Processes and Suggestions for Improvements*, Santa Monica, Calif.: RAND Corporation, RR-471-AF, 2014. As of October 19, 2017:
https://www.rand.org/pubs/research_reports/RR471.html

Stevenson, J. M., D. R. Greenhorn, J. T. Bryant, J. M. Deakin, and J. T. Smith, "Selection Test Fairness and the Incremental Lifting Machine," *Applied Ergonomics*, Vol. 27, 1996, pp. 45–52.

Teves, Marilyn A., James E. Wright, and James A. Vogel, *Performance on Selected Candidate Screening Test Procedures Before and After Army Basic and Advanced Individual Training*, Ft. Detrick, Md.: Army Medical Research and Development Command, June 1985.

Truxillo, D. M., L. M. Donahue, and J. L. Sulzer, "Setting Cutoff Scores for Personnel Selection Tests: Issues, Illustrations, and Recommendations," *Human Performance*, Vol. 9, No. 3, 1996, pp. 275–295.

U.S. Department of Defense, *Women in the Service Implementation Plan*, memorandum from the Chairman of the Joint Chiefs of Staff, January 9, 2013.

U.S. General Accounting Office, *Physically Demanding Jobs: Services Have Little Data on Ability of Personnel to Perform*, report to the Chairman, Subcommittee on Military Personnel, Committee on National Security, House of Representatives, GAO/NSIAD-96-169, July 1996.

Williford, H. N., W. J. Duey, N. S. Olson, R. Howard, and N. Wang, "Relationship Between Fire Fighting Suppression Tasks and Physical Fitness," *Ergonomics*, Vol. 42, No. 9, 1999, pp. 1179–1186.