Project Report

Development and Validation of Physical Performance Tests for Selection into USAF Occupational Specialties

RAND Project AIR FORCE and Human Resources Research Organization (HumRRO)

PR-2098-1-AF
September, 2015
Prepared for the United States Air Force

NOT CLEARED FOR OPEN PUBLICATION

This document has not been formally reviewed, edited, or cleared for public release. It may not be cited, quoted, reproduced or transmitted without the permission of the RAND Corporation. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors. RAND® is a registered trademark.

Summary and Key Findings

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women from assignment to units and positions whose primary mission is to engage in direct combat on the ground, was rescinded.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 integration process, Joint Staff guidance and federal laws require that eligibility and occupational standards for all occupations reflect job tasks[1]. To be in compliance with this mandate, RAND, Human Resources Research Organization (HumRRO), and the Air Force developed and executed a methodology to examine the validity of the Strength Aptitude Test (SAT), which is a physical strength test used at the Military Entrance Processing Stations (MEPS) to classify recruits into Air Force Specialties (AFSs). As part of the study to examine the validity of the SAT, other physical fitness tests were examined to identify whether alternative tests would have higher validity than the SAT and/or reduced adverse impact against women.

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 to be in compliance with federal laws, this study was designed to address the following questions:

1. What are the physical requirements to perform in different AFSs?
2. How can physical performance on job-relevant tasks be measured?
3. Which physical fitness tests, including the SAT, are valid indicators of an individual'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)?

Based on these questions, a criterion-related validity approach was used for this study. A criterion-related validation strategy demonstrates the statistical relationship between the physical fitness tests and job performance. One approach to criterion-related validity is concurrent validation, which involves measuring performance on the fitness tests (i.e., predictors of performance) and the measures of job performance (i.e., the criterion) with the same group of

[1] DOD, memorandum from the Chairman of the Joint Chiefs of Staff, *Women in the Service Implementation Plan* (Jan. 9, 2013); Pub. L. No. 113-219 § 524 (2014).

individuals concurrently. This study used a concurrent validation design in which the physical performance of USAF personnel was assessed using the physical fitness tests as measures of physical abilities and work samples as measures of physical job performance.

Key Findings

This study integrated scientific and professional best practices to evaluate the validity evidence for the SAT. Not only did the study find strong evidence for the validity of the SAT, but it also identified several other tests that could be used in some combination to further strengthen the validity of entry-level physical fitness testing. These additional tests should be further evaluated to balance the relative gains of increased validity with the costs of additional testing at the MEPs. In conclusion, 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.

What Are The Physical Requirements For AFSs?

Professional best practice for identifying job requirements is to conduct a job analysis. Job analysis involves systematically gathering and evaluating information about work activities, worker attributes, and the work context to make inferences about job requirements. The Air Force's Occupational Analysis Division conducts job analyses on each enlisted occupational specialty on a three-year cycle (Sims et al., 2014). The results of these job analyses are compiled into Occupational Analysis Reports (OARs), which provide detailed information on the airmen such as skill level, job tasks performed, and the percentage of airmen performing each task. 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 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.

Job Analysis and Test Development

Considering available time and resources, RAND and HumRRO determined that a sampling strategy was needed to ensure representation of the range of physical requirements across AFSs. The first step was to review AFSs (n = 30) with a current SAT requirement of 70 lbs. or higher. Although these AFSs perform the most physically demanding tasks in the AF, other AFSs with a SAT requirement less than 70 could have slightly different physical requirements worth considering. Therefore, HumRRO also reviewed 87 AFSs in the less than 70 lb. category for inclusion in the study.

Identification of AFSs for Inclusion in the Study

Based on a review of USAF job materials that include OARs, Career Field Education and Training Plans, and PowerPoint briefings, 21 AFSs with varying levels of physical demands were used in the job analysis. The criteria used in the selection of AFSs are listed below.

1. Weight lifted
2. Movement Categories
    a. Lift/carry
    b. Push/pull
    c. Climb
    d. Walk/stand

   e. Hold
   f. Shovel/dig
   g. Pound/hammer
3. Prior knowledge of job components
4. Type of career field
5. Shred level

The goal was to select AFSs that were representative of varying levels of physical demand. Use of AFSs with varying demands allowed for differences in the ergonomic parameters associated with the movement categories associated with task performance. For example, the Hold movement category includes ergonomic parameters such as object weight, height held at, and size of object. Thus, AFSs with greater physical demands would most likely hold heavier objects than ones with lower demands. This approach ensured that input of ergonomic and other data into the new prediction model would accurately classify the current and new AFSs in terms of physical demand.

Job Analysis Data Collection

As previously mentioned, existing OARs do not provide specific details about the physical requirements for each AFS. Therefore, RAND and HumRRO developed a set of job analysis questions to use for interviewing individuals who are knowledgeable about an AFS's physical demands. These questions were integrated into a comprehensive Movement Classification Questionnaire (MCQ) that can be used to identify all job-relevant movement categories associated with physically demanding tasks performed in the AF. HumRRO identified movement categories that possess physical demand in terms of muscular strength, muscular endurance, aerobic capacity, and anaerobic power. The movement categories identified for the MCQ are listed below.

- Lift
- Carry
- Push/Pull
- Climb
- Stand
- Non-Standing Position (e.g., kneel)
- Walk
- Run
- Crawl
- Hold
- Shovel
- Dig

- Pound
- Swim
- Operate Powered Hand-Held Tools
- Operate Non-Powered Hand Tools

Within each of these movement categories, the MCQ contained questions to capture information on the ergonomic parameters that affect the physical demand of each movement category. These questions were identified by reviewing ergonomic literature and other research that addressed one or more of the movement categories. For example, the ergonomic parameters that affect the physical demand of a "Lift" include items such as object weight, percent of time lifted with assistance, number of items lifted at one time, number of items lifted in task, height lifted from, and height lifted to.

After further review and pilot testing, the MCQ was administered by RAND personnel via interviews and by HumRRO staff in focus group meetings. Prior to the interviews, interviewees were provided with a list of the 10 most physically demanding tasks within their AFSs, which were selected by career field managers or similar subject matter experts familiar with the physical demands of the job. The format for the interviews consisted of three steps.

1. Obtain demographic information from the interviewees (e.g., current rank, age, number of times deployed).
2. Determine variability in task performance and comprehensiveness of the 10 tasks.
    a. 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)?
    b. Are there any other physically demanding tasks not included in the current list?
3. Gather detailed information on the physical demands for each of the 10 tasks, including the movement categories that apply (e.g., lifting) and the parameters that could affect the physical demand of that task (e.g., weight of object). Table 1 provides an example of the questions used for one of the movement categories.

Table 1. Example Information Obtained in Job Analysis Data Collection Instrument

Item #1: What is being lifted?
Item #1: Weight maximum (lb)
Item #1: Weight minimum (lb)
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

In addition to conducting interviews with incumbents, HumRRO staff observed incumbents performing job tasks at Air Force Bases (AFBs). The observations focused on the 10 tasks in the MCQ for an AFS. Due to the nature of some of the tasks (e.g., only performed in emergency situations, not performed on a daily basis, only performed during deployment), it was not possible to observe all tasks. If a task could not be observed, the incumbents showed the types of equipment used and simulated performing parts of the tasks.

Physical Abilities Required

The job analysis results showed that a variety of physical abilities were related to the 16 movement categories in the MCQ. Table 2 provides a summary of the movement categories and their associated physical abilities. These results show that each of the physical abilities was linked to nine or more of the 16 movement categories, except for manual dexterity. Muscular strength and muscular endurance were linked to 16 and 15 of the movement categories, respectively.

Table 2. Basic Physical Abilities Linked to Movement Categories

	Physical Ability							
Movement Category	**Muscular Strength**	**Muscular Endurance**	**Aerobic Capacity**	**Flexibility**	**Equilibrium**	**Anaerobic Power**	**Coordination**	**Manual Dexterity**
Lift	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Carry	Yes	Yes	Yes	No	Yes	Yes	Yes	No
Push/ Pull	Yes	Yes	Yes	No	Yes	Yes	Yes	No
Climb	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Stand	Yes	Yes	No	No	Yes	No	No	No
Non-standing position (e.g., kneel, squat)	Yes	Yes	No	Yes	Yes	No	Yes	No
Walk	Yes	Yes	Yes	No	Yes	No	No	No
Run	Yes	Yes	Yes	No	Yes	Yes	Yes	No
Crawl	Yes	No	No	Yes	No	No	No	No
Hold	Yes	Yes	No	No	No	Yes	Yes	No
Shovel	Yes	Yes	Yes	Yes	No	Yes	No	No

Identification of Physical Fitness Tests

Past validation research was reviewed to identify physical tests that assessed the relevant abilities that were significantly related to measures of job performance (e.g., Blakley, Quinones, Crawford, & Jago, 1994; Gebhardt & Baker, 2010a; Rayson et al., 2000). This review resulted in the identification of a variety of physical tests. It also showed that tests of flexibility (e.g., Sit & Reach, Twist & Touch) were rarely related to job performance (e.g., Gebhardt & Baker, 2010a). Thus, measures of flexibility were not included in the test development. Additionally, manual dexterity was eliminated from further consideration because it was only identified for two of the 16 movement categories.

Following this review, a total of 19 tests were identified as possible candidates for the validation study. However, several tests were further discarded due to time constraints in the time allotted to test research participants. Similarly, other tests were eliminated due to potential space and/or time restrictions at the MEPS, where future testing would take place. The final set of 9 fitness tests are briefly described below.

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 & 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 of 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, Herrin, Keyserling, & Garg, 1977; Gebhardt & Baker, 2010b).

Handgrip

The Handgrip test measures grip strength. Three trials are given for both the dominant and non-dominant hands. A mean is calculated for the dominant and non-dominant 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=.63) (Gebhardt & Baker, 2010b).

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 & 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 & 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 & 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 take-off and landing is used, with swinging of the arms and flexing of the knees to provide forward propulsion (Koch, O'Bryant, Stone, Sanborn, Proulx, Hruby, J., 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, Santa Maria, Davis, & Schwartz, 1978), as well as highly related ($r = 0.81 – 0.84$) to carrying objects and stretchers (Bilzon, Scarpello, Bilzon, & Allsopp, 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 & Baker, 2010b).

Strength Aptitude Test (SAT)

The SAT was developed by the USAF to assess the strength of recruits. The SAT is performed using an incremental lifting machine (ILM) with handles that rotate to accommodate the change in hand position when lifting from six inches above the ground to a height of six feet. The initial weight lifted is 40 lb. If an individual is successful at this weight, 10 lb. are added to the ILM and a lift is performed with 50 lb. The weight is increased 10 lb. until the individual is no longer able to lift to six feet or attains a score of 110 lb., which is 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, & Thune, 2006; Knapik et al., 2004; Myers et al., 1984; Teeves, Wright, & Vogel, 1985).

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 lb. If the lift of 110 lb. was successful, 10 lb. were 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 lb. One hundred and ninety pounds was selected as the SAT's endpoint because the ILM's maximum was 190 lb. 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 lb. 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 increase women's scores on the SAT.

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

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 work samples. 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, work samples are a more direct, objective measure of physical performance because they approximate the physical demands of job-relevant tasks.

Development of Job Performance Measures

HumRRO reviewed the job analysis results from the MCQ to determine movement categories required by most of the 21 specialties included in the job analysis. This review found that many of the 10 physically demanding tasks within an AFS contained the same movement categories across the 21 AFSs. Furthermore, the movement categories of Lift, Carry, Push/Pull, Climb, Stand, Non-Stand/Kneel, Hold, and Operate Non-Power Tools are linked to the 75 percent or greater of the AFSs (19 to 21), while Walk, Crawl, Pound, and Operate Power Tools are linked to approximately 50 percent of the AFSs. The remaining movement categories (Run, Shovel, Dig, Swim) were linked to less than 50 percent of the AFSs.

Movement categories linked to the 75 percent or greater of the AFSs were retained for work sample development. These movement categories are Lift, Carry, Push/Pull, Climb, Stand, Non-Stand/Kneel, Hold, Operate Non-Power Tools. The results yielded eight potential work samples. Following further review of the work samples and to stay within the four-hour time constraint for testing, the number of work samples was reduced to four. These four work samples are described below.

- The Lift/Carry work sample consisted of lifting equipment associated with multiple AFSs 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 process was repeated and all objects were moved back to their starting positions. This completed one cycle of the Lift/Carry work sample. A total of two cycles were completed in this work sample.
- The Push/Pull work sample involved pushing portable lights, followed by 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 were completed.

- The Climb work sample consisted of climbing an extension ladder to a height of 9 feet while wearing a 30-lb vest. When an airman's feet reached the 9th rung (9 feet), the airman moved a small simulated tool box (14 lb) 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 work sample 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 the work sample.
- The Hold work sample consisted of three cycles. 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 10-second rest period. The third cycle used the same protocol but was performed from a squatting position.

Table 3 provides a summary of the relevance of each work sample to each AFS included in the job analysis. The table shows that the four work samples were relevant to most AFSs.

In addition to pilot testing each of the work samples, data were collected from airmen to determine the test-retest reliability of the work samples. Test-retest reliability is the degree to which participants' scores remain relatively consistent over repeated administrations of the same measure (Crocker & Algina, 1986). Test-retest reliability coefficients are computed by administering a test to a group of participants at two different points in time and then correlating their two sets of scores. The obtained correlation coefficient provides an indication of 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 work samples were high and ranged from 0.77 (Hold) to 0.93 (Lift/Carry). The high correlations indicate that these four work samples are consistent measures of physical task performance (Dancey & Reidy, 2004).

Table 3. Linkage of Work Samples to AFSs

			Work Samples by Movement Category			
Job#	**AFSC**	**AFS 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 & 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

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

To evaluate the validity of the fitness tests, work sample scores were first combined into a single composite score to provide an overall measure of physical job performance. That is, the composite work sample score was the sum of standardized scores on the four work samples: a) Climb and Carry, b) Hold, c) Lift and Carry, and d) Push and Pull. After composite work sample scores were computed, scores on each fitness test were correlated with scores on the work sample composite. After reviewing the validity of the individual tests, steps were taken to evaluate whether a combination of tests (i.e., a test battery) could strengthen the prediction of performance.

Validity of Individual Physical Fitness Tests

The results indicated that all of the fitness tests in the study were significantly correlated with the work sample composite (see Table 4)[2]. Since all of the tests were significantly correlated with the composite, it was determined that any of these tests could be considered in the final test battery.

In addition to its significant correlation with the composite, the SAT had significant correlations with all four individual work samples. These significant correlations demonstrate that the SAT is predictive of physical job performance, which indicates that higher SAT scores are associated with better work sample performance. These results provide strong evidence for the validity of the SAT test.

[2] Correlations are positive for the Hold work sample because longer times for holding objects indicate better performance. Conversely, shorter times on Climb and Carry, Lift and Carry, and Push and Pull work samples indicate better performance.

Table 4. Correlations Between Fitness Tests and Work Samples

Fitness Test	Climb & Carry Final Time	Hold All Cycles Total Time Held	Lift/Carry Final Time	Push/Pull Final Time	Composite (All Work Samples)
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

Identifying a Potential Test Battery

A series of statistical analyses were conducted to identify whether combining tests would result in a better prediction of work sample performance. Overall, the results indicated several possible combinations of tests that could strengthen validity. Each of the different models resulted in a slightly different combination of four tests, and the test combinations have different advantages and disadvantages. Some combinations would require the purchase of relatively expensive equipment (i.e., arm ergometer), and some combinations result in higher levels of adverse impact against women than other combinations and do not provide gains in validity. One model meets several important objectives: a) increases validity, b) minimizes equipment costs, and c) minimizes adverse impact against women. This model, which was identified using a regression analysis, shows that four tests – Handgrip (Best 2 Trials), SAT, Sit-Ups, and Step Test VO2 – resulted in a significant multiple R of 0.81. The squared multiple R of 0.81 indicated that approximately 81 percent of the variance in the job performance (measured by the composite work sample) was accounted for by the four fitness tests. Moreover, this model provides an 8 percent increase in validity over using the SAT by itself. Using a test battery with stronger validity will result in a more accurate classification by assigning airmen into occupational specialties for which they have a high probability of successfully performing the physical tasks required in their specialties.

Do The Fitness Tests Predict Physical Performance Equally Well For Different Subgroups?

The Uniform Guidelines on Employee Selection Procedures (Equal Employment Opportunity Commission, 1978) and the Standards for Educational and Psychological Testing (American Psychological Association, 1999) recommend that when using a test battery to make selection decisions, users investigate differences in criterion-related validity evidence for various demographic subgroups (e.g., sex, race and national origin, age), where it is technically feasible. This examination ensures that the test battery is fair for the various subgroups.

To evaluate the fairness of the test battery, a differential prediction approach was used. The term *differential prediction* refers to a significant difference in the regression equations for two groups, which is indicated by differences in the slopes, the intercepts, or both. Differences in intercepts indicate members of one group tend to obtain lower criterion scores than members of another group who have the same predictor test score. Differences in slopes indicate the regression lines intersect at some point, so there is overprediction of the criterion scores of one group compared to the other below the intersection, and underprediction above the intersection.

Analyses for test fairness were conducted not only for the SAT but for each of the test battery models that were considered. The results from each of these analyses were taken into consideration when making final recommendations for tests to use.

Conclusion

This study integrated scientific and professional best practices to evaluate the validity evidence for the SAT. Not only did the study find strong evidence for the validity of the SAT, but it also identified several other tests that could be used in some combination to further strengthen the validity of entry-level physical fitness testing. These additional tests should be further evaluated to balance the relative gains of increased validity with the costs of additional testing at the MEPs. In conclusion, 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.

References

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education (Joint Committee), *Standards for Educational and Psychological Testing (2nd edition)*, Washington, D.C.: American Psychological Association, 1999.

Ayoub, M. M., Jiang, B. C., Smith, J. L., Selan, J. L. and McDaniel, J. W., "Establishing a physical criteria for assigning personnel to U.S. Air Force jobs," *American Industrial Hygiene Association Journal*, Vol. 48, 1987, pp. 464- 470.

Baker, T. A., and Gebhardt, D. L. Chapter 13: The assessment of physical capabilities in the workplace. In N. Schmitt (Ed.), *The Oxford Handbook of Personnel Assessment and Selection*, New York, N.Y.: Oxford University Press, Inc., 2012.

Bilzon, J. L., Scarpello, E. G., Bilzon, E., and Allsopp, A. J., Generic Task-related occupational requirement for Royal Naval (RN) personnel, In P. McCabe (Ed.), *Contemporary Ergonomics*, United Kingdom: Institute of Naval Medicine & Taylor Francis Ltd.,2003.

Blakley, B. R., Quiñones, M. A., Crawford, M., and Jago, I. "The validity of isometric strength tests," *Personnel Psychology*, Vol. 47, No. 2, 1994, pp. 247-274.

Chaffin, D. B., Herrin, G. D., Keyserling, W. M., and Garg, A., "A method for evaluating the biomechanical stresses resulting from manual materials handling jobs," *American Industrial Hygiene, Association Journal*, Vol. 38, 1977, pp. 662-675.

Crocker, L. A, and Algina, J., *Introduction to Classical and Modern Test Theory*. Fort Worth, Tex.: Harcourt, 1986.

Dancey, C., & Reidy, J., *Statistics without Maths for Psychology: Using SPSS for Windows*, London: Prentice Hall, 2004.

Dotson, C., Santa Maria, D., Davis, P., and Schwartz, R., *Development of a job-related physical performance test for firefighters*. College Park, MD: University of Maryland and U.S. Department of Commerce, 1978.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice, *Uniform Guidelines on Employee Selection Procedures*, Washington, D.C.: Bureau of National Affairs, Inc., 1978.

Gebhardt, D. L., and Baker, T. A., Physical performance tests, In J. Farr & N. Tippins (Eds.), *Handbook of Employee Selection*, New York, N.Y.: Routledge, Chapter 13, 2010a.

Gebhardt, D. L., & Baker, T. A., Physical Performance. In J. Scott & D. Reynolds (Eds.), *Handbook of Work Assessment*, San Francisco, Calif.: Jossey-Bass, Chapter 7, 2010b.

Gebhardt, D. L., Baker, T. A., & Thune, A. (2006). Development and validation of physical, cognitive, and personality assessments for selectors and delivery drivers, Beltsville, MD: Human Performance Systems, Inc.

Koch, A. J., O'Bryant, H. S., Stone, M. E., Sanborn, K., Proulx, C., Hruby, J., Shannonhouse, E., Boros, R., and Stone. M. H., "Effect of warm-up on the standing broad jump in trained and untrained men and women,". *The Journal of Strength & Conditioning Research*, Vol. 17, 2003, pp. 710-714.

Knapik, J. J., Darakjy, S., Scott, S., Hauret, K. G., Canada, S., Marin, R., Palkoska, F., VanCamp, S., Piskator, E., Rieger, W., and Jones, B. H., *Evaluation of Two Army Fitness Programs: The TRADOC Standardized Physical Training Program for Basic Combat Training and the Fitness Assessment Program*, Aberdeen Proving Ground, Md.: U.S. Army Center for Health Promotion and Preventive Medicine, 12-HF-5772B-04, 2004.

Myers, D. C., Gebhardt, D. L., Crump, C. E. and Fleishman, E. A., *Validation of the Military Entrance Physical Strength Capacity Test*, Technical Report 610 Natick, U.S. Army Research Institute of Environmental Medicine, 1984.

Rayson, M., Holliman, D. and Belyavin A,. "Development of physical selection procedures for the British Army. Phase 2: Relationship between physical performance tests and criterion tasks," *Ergonomics*, Vol. 43, No.1, 2000, pp. 73-105.

Sims, C. S., Hardison, C. M., Keller, K. M., & Robyn, A., *Air Force personnel research: Recommendations for improved alignment*, Santa Monica, Calif.: RAND Corporation, RR-814-AF, 2014.

Teeves, M. A., Wright, J. E., and Vogel, J. A., *Performance on selected candidate screening tests and procedures before and after army basic and advanced individual training* (Report No. T13/85), Natick, Mass.: U.S. Army Research Institute of Environmental Medicine, 1985.