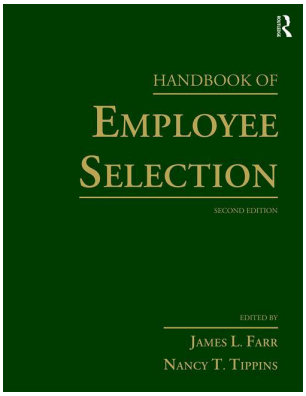

Handbook of Employee Selection

James L. Farr, Nancy T. Tippins, Walter C. Borman, David Chan, Michael D. Coovert, Rick Jacobs, P. Richard Jeanneret, Jerard F. Kehoe, Filip Lievens, S. Morton McPhail, Kevin R. Murphy, Robert E. Ployhart, Elaine D. Pulakos, Douglas H. Reynolds, Ann Marie Ryan, Neal Schmitt, Benjamin Schneider

Physical Performance Tests

Publication details
https://test.routledgehandbooks.com/doi/10.4324/9781315690193-12
Deborah L. Gebhardt, Todd A. Baker
Published online on: 22 Mar 2017

12

PHYSICAL PERFORMANCE TESTS

DEBORAH L. GEBHARDT AND TODD A. BAKER

The Bureau of Labor Statistics (2011) reported that 28% of the workforce in the United States performs physically demanding jobs that involve construction, machinery installation and repair, public safety, and other professions. In many instances, physically demanding jobs are the highest paying jobs in a geographic location. The importance of ensuring access to these jobs for all applicants was underscored in the Department of Defense (DoD) decision in 2015 to open all military occupational specialties (MOS) to men and women (Carter, 2015).

Assessment of physical performance has a historical base in the fields of exercise science, medical, psychology, and military and encompasses the academic areas of physiology, biomechanics, industrial engineering, applied psychology, and medicine. These multidisciplinary aspects led to the use of physical testing in occupational settings. Although our society has become more computer-driven, there are many arduous jobs in the public, private, and military sectors. The warehouse, manufacturing, long-shore, telecommunications, railroad, airline, electric, and natural gas industries contain many arduous jobs (Gebhardt & Baker, in press).

Organizations use physical performance tests for applicant selection, retention of incumbents, and evaluation of physical fitness levels. Although physical testing is common in the selection setting, some organizations evaluate incumbent personnel at specified intervals (e.g., annually) to determine if they can perform the physical aspects of the job. A few public safety agencies require annual physical qualification (e.g., Nuclear Regulatory Commission, state police agencies) with employment consequences ranging from remedial training, denial of promotion and bonus payments, and job suspension, to job loss (Gebhardt & Baker, 2006).

JOB ANALYSIS FOR ARDUOUS JOBS

Similar to all assessments used when making employment decisions, physical tests must be supported by a detailed job analysis. In the physical area, it is important to consider all underlying parameters (e.g., environment, protective equipment) that affect job performance. It would be unrealistic to consider police officer job tasks without including the weight of the equipment worn (e.g., bulletproof vest, weapon, ammunition, radio, handcuffs). As with all job analyses, identification of essential tasks and abilities is critical to defining job requirements. The job analysis—whether physical, cognitive, or psychomotor—involves three steps: job observations and interviews, identification of essential tasks, and identification of ergonomic and environmental conditions.

Site Visits and Essential Task Identification

Job site visits involve interviews and observation of incumbents performing job tasks. During the observations, researchers identify job tasks and gather data related to postures, motions, and ergonomic parameters associated with the tasks. Interviews with incumbents and supervisors provide lists of tasks and details related to task performance (e.g., equipment weights). When the intent of the job analysis is to develop and validate physical performance tests, the task statements should be specific in nature to allow for identification of the frequency of specific types of physical activities (e.g., lifting different types and weights of objects). Following these initial steps, incumbents and supervisors use rating scales such as frequency of performance and importance to the job to identify essential job tasks. Use of other scales such as *physical effort* assists in initially obtaining an overview of the physical demand of job tasks (Fleishman, Gebhardt, & Hogan, 1986). Tasks with mean physical effort ratings equal to or above a specified value (e.g., 4 on a 7-point scale) contain moderate or higher physical demand. The *expected to perform* scale, used for public safety jobs, identifies rarely performed tasks that are critical to successful job performance (e.g., discharging firearms, carrying victims from burning structures).

Past research found that job incumbents and supervisors provide reliable task ratings (Hogan, 1991a). However, incumbents provide better frequency and time spent ratings, in most instances, because they perform the tasks. If supervisors are used, first-line supervisors with previous job experience are most appropriate.

Researchers have used numerous decision rules or algorithms (frequency, importance, and time-spent combinations) to identify essential tasks from task ratings. There is not one specific algorithm associated with physical-oriented job analysis. Selection of the most appropriate algorithm depends upon the nature and mission of the job. For instance, jobs in which most tasks are performed frequently (e.g., assembly line) may require a larger weighting for frequency than importance. Jobs that include highly important tasks that are infrequently performed (e.g., discharge firearm) may use an algorithm in which there is a separate importance or frequency mean cutoff to identify essential tasks. Thus, there may be need for a combination of algorithms to define the essential physical tasks.

Ergonomic/Biomechanical/Physiological Analysis

Ergonomic, physiological, and biomechanical data, used separately or in combination, provide direct measures to quantify physical job demands. The methodologies range from simple measures such as the distance a worker walks, to sophisticated measures involving oxygen consumption, mathematical modeling, and use of archival engineering data. Simple ergonomic measures such as weights of objects, distances objects carried, and heights lifted to and from are appropriate for most jobs. To measure the force required to move objects, researchers use a load-cell device that records force production.

To quantify actual job task demand, researchers use basic physiological and biomechanical data-gathering methodologies. The type of data gathered is dependent upon the essential tasks and physical demands of the job. The data can be gathered using a variety of equipment (e.g., heart rate monitor, accelerometer, oxygen/gas sensor, mathematical modeling). Heart rate monitors can attain a basic estimate of the physiological workload for jobs requiring task performance at medium to high intensities for extended timeframes (e.g., order filler, firefighter). The monitor captures the individual's heart rate during task performance, while the researchers calculate the heart rate response and the percentage of maximum heart rate at which the individual was working. For example, if a 30-year-old male with a maximum heart rate of 190 beats per minute (bpm) (220) – age (30) = 190 bpm) is working at an average heart rate of 142.5 bpm, then he is working at 75% of his maximum (142.5/190 = 0.75). The American College of Sports Medicine (ACSM) classified the intensity of physical activity in terms of the percentage of maximum heart rate (Pescatello, Arena, Riebe, & Thompson, 2014). Table 12.1 lists the ACSM intensities, which range from very light to maximum (Pescatello et al., 2014). Gebhardt,

TABLE 12.1
ACSM's Categories of Physical Activity Intensity

Intensity	Percentage of Maximum Heart Rate
Very light	<35
Light	35–54
Moderate	55–69
Hard	70–89
Very hard	≥90
Maximum	100

Baker, and Thune (2006) found that workers in an order filler job had heart rates of 71–81% of maximum across a 3- to 4-hour timeframe, thus placing the job in the "hard" intensity level. Use of this information and other data helped determine an estimate of the aerobic capacity ($VO_{2submax}$) needed to perform job tasks.

Past research indicated that to sustain arduous work for an 8-hour period, one should not exceed 40–50% of maximum aerobic capacity ($VO_{2\,max}$) (Astrand, Rodahl, Dahl, & Stromme, 2003; McArdle, Katch, & Katch, 2015). Direct measures of oxygen uptake have been performed on jobs ranging from light industry (e.g., manual materials handling) to firefighting and military jobs (Bilzon, Scarpello, Smith, Ravenhill, & Rayson, 2001; Sothmann, Gebhardt, Baker, Kastello, & Sheppard, 2004). The VO_2 requirements for shipboard, urban, and forest firefighting ranged from 33.5 to 45.0 milliliters of oxygen/kilogram of body weight/minute ($mL \cdot kg^{-1} \cdot min^{-1}$) (Bilzon et al., 2001; Gledhill & Jamnik, 1992; Sothmann et al., 2004). The oxygen consumption required to perform an emergency response involving restraining and subduing an individual ranged from 38.5 to 39.5 $mL \cdot kg^{-1} \cdot min^{-1}$, respectively (Jamnik, Thomas, Burr, & Gledhill, 2010). This type of physiological data, along with heart rate data, is helpful in determining whether an aerobic capacity selection test would be beneficial and in establishing job-related passing scores.

Use of biomechanical data encompasses the use of physics principles to define human movement. If the force and other parameters (e.g., torque) are not available through direct measurement (e.g., load cell to determine force to move object), researchers videotape task movements and calculate the forces, torques, and acceleration components. Biomechanical models (mathematical) provide an avenue to assess physical demand. For instance, researchers developed a model to identify the forces required by paramedics to lift a patient-loaded stretcher into an ambulance based on ergonomic parameters of the stretcher (e.g., length, weight) and the height and weight of patients (Gebhardt & Crump, 1984). This model indicated that a force of 152 pounds was required to lift the head end of a stretcher carrying a 200-pound patient. Another method, motion analysis, requires videotaping workers performing a job task. The motions are captured using optical sensors placed at the subjects' joints (e.g., elbow). These data are mathematically transformed to provide indications of the forces incurred at specific anatomical locations (e.g., knee, hip), which yield an indication of the forces required to complete the task.

In summary, ergonomic, biomechanical, and physiological data provide information important to defining the physical demand of a job. These data also form the basis for developing predictor tests and criterion measures and setting passing scores. Some of these data are available in the literature, but others must be obtained through direct measurement at the job site or other location.

Identification of Environmental Conditions

Environmental working conditions (e.g., heat, surface conditions) play an integral part in the performance of physical tasks. Researchers use job analysis questionnaires, incumbent focus

groups, standard operating procedures, and past weather history to obtain environmental condition information. Arduous work performed in high temperatures (e.g., 90°F or greater) and/or occlusive clothing increases the physical demand and time required to complete job tasks. For example, nuclear power plant workers wear occlusive clothing to protect them from the radiation. This clothing increases the workers' core temperature, which causes excessive sweating and reduces the workers' ability to perform job tasks. Research showed that women do not dissipate heat as well as men when performing arduous tasks in hot environments (Epstein, Yanovich, Moran, & Heled, 2013). Conversely, in cold environments, when matched by body size, men and women lose heat at similar rates, but women perform physical and cognitive tasks better than men at lower body temperatures (Solianik, Skurvydas, Mickevičienė, & Brazaitis, 2014; Tikuisis, Jacobs, Moroz, Vallerand, & Martineau, 2000). Other research found that individuals with higher aerobic capacity are more readily able to adjust to a heated environment (Astrand et al., 2003; Pandolf, Burse, & Goldman, 1977). Thus, defining the demands of the essential tasks may assist in designing the testing procedures, as well as criterion measures used in validation studies.

Identification of Required Physical Abilities

Identification of the physical abilities required for a position provides an overview of the job demands. Past research defined physical abilities in several contexts. Listed as follows are the physiological definitions (Astrand et al., 2003; McArdle et al., 2015):

1. Muscular strength is the ability to exert force to lift, push, pull, or hold objects.
2. Muscular endurance is the ability to exert force continuously over moderate to long timeframes.
3. Aerobic capacity is the ability of the respiratory and cardiovascular systems to provide oxygen to the body systems for medium- to high-intensity tasks performed over a moderate timeframe.
4. Anaerobic power is the ability to complete high-intensity, short-duration (e.g., 5–90 seconds) tasks using stored energy (e.g., adenosine triphosphate).
5. Flexibility involves the range of motion at the joints (e.g., knee, shoulders) to bend, stoop, rotate, and reach in all directions with the arms and legs.
6. Equilibrium is the ability to maintain the center of gravity over the base of support (e.g., feet) when outside forces (e.g., gravity, slipping on ice) occur.
7. Coordination is the ability to integrate sight, hearing, and other neuro-sensory cues to perform motor activities (e.g., change of direction) in an accurate sequential pattern.

Other research identified different factor structures for classifying physical abilities. One taxonomy was similar to physiological abilities and included nine physical abilities (e.g., static strength, dynamic strength, coordination, equilibrium), which are included in the O*NET (Fleishman & Quaintance, 1984; Fleishman, 1964). Another study found three components: muscular strength, endurance, and movement quality (Hogan, 1991b). A subsequent study using equal samples of men and women found a six-factor structure best described physical performance (Myers, Gebhardt, Crump, & Fleishman, 1993). Guion (1998) compared several physical ability classifications and grouped muscular strength, muscular endurance, and muscular power (anaerobic power) into a muscular strength factor and the remaining factors into a movement quality factor. However, use of a single strength factor did not correspond to the physiological components that underlie performance of different types of physical tasks. For example, it takes 5–10 minutes to complete 300 turns when closing large wheel valves, thus requiring muscular endurance. Using a single strength factor would not adequately define the physiological demand of this task and could lead to use of the wrong test for applicant selection. Although each of these structures has scientific merit, a combination of these studies provides a framework for identifying physical requirements in the work setting. These abilities are muscular strength, muscular endurance, aerobic capacity, anaerobic power, flexibility, and equilibrium, along with a coordination factor.

Performance of physical tasks requires varying levels of the different physical abilities. Muscular strength may be as minimal as lifting a spoon or as high as lifting 90-pound cement bags. Similarly, energy expenditure may be primarily anaerobic (e.g., drag a victim 50 feet) or aerobic

(e.g., fill eight warehouse orders totaling 8,900 pounds) depending on the duration and intensity of the activity. When identifying the relevant physical abilities for a position, it is important to gather information related to the level of the physical abilities needed to complete essential job tasks. Two methods provide this information. One involves direct measurement of the job task(s) as described above.

The second method uses physical ability-rating scales with behavioral anchors targeted at work behaviors or physical activities (e.g., climb 20-foot ladder, jog 3 miles) (Fleishman & Quaintance, 1984; Gebhardt, 1984). Incumbents, supervisors, or job analysts rate essential job tasks on each scale to identify the amount of the ability needed to complete the task. The consolidation of the ratings produces a profile of the physical demand of a job. This approach allows for comparison of multiple jobs and assists in the selection or design of testing procedures for relevant abilities. Regardless of the method used to determine the abilities related to job performance, identification of the job-related abilities and their magnitudes provides a link between the essential job tasks and the physical tests designed for use in selection and retention settings.

PHYSICAL PERFORMANCE TESTS

There are two types of physical tests: basic ability and job simulation assessments. Basic ability tests measure a single ability or construct (e.g., muscular strength, flexibility) and typically do not resemble job tasks. These tests assess the physical abilities required for performance of essential job tasks. Use of basic ability tests allows for assessment of multiple jobs that require the same abilities. Examinees typically perform simple movements (e.g., elbow flexion, stepping onto a platform at a specified cadence) in a basic ability test, thus resulting in a low risk of injury for applicants. Several overviews of basic ability tests are located in other reviews (Baker & Gebhardt, 2012; Landy et al., 1992; Reilly, Gebhardt, Billing, Greeves, & Sharp, 2015; Tipton, Milligan, & Reilly, 2013).

Muscular strength tests fall into three categories: isometric, isotonic, and isokinetic. *Isometric* or static strength tests require exerting a maximum force without movement at the joint (e.g., elbow). In this type of test, a muscle group generates force, but the length of the muscles remains unchanged (Astrand et al., 2003; McArdle et al., 2015). The arm lift test is an example of an isometric test, and requires holding a bar with the elbows flexed to 90 degrees and exerting an upward force (Chaffin, Herrin, Keyserling, & Foulke, 1977). The score is the force generated. Isometric shoulder, arm, torso, and leg strength tests have been used extensively in selection settings and were valid predictors of job performance ($r = .39$ to .63) (Blakley, Quinones, Crawford, & Jago, 1994; Gebhardt, Baker, & Sheppard, 1998; Jackson & Sekula, 1999).

Isotonic tests measure the force generated by a muscle group through a range of motion at one or multiple joints (e.g., hip, knee) (Astrand et al., 2003; McArdle et al., 2015). Tests such as one repetition bench press or a dynamic lift to a specified height are examples of isotonic tests. Isotonic tests were significant predictors of job performance in public safety and industrial jobs (Davis, Dotson, & Santa Maria, 1982; Gebhardt & Crump, 1984).

Isokinetic testing assesses the force produced through a specified range of motion at the shoulder, back, and knee joints. The equipment incorporates a force-recording device (load cell) and computer software, which controls the speed (degrees/second) at which a subject can perform maximal flexion and extension movements. The measurement unit for the force generated by a subject is torque (τ), a vector quality that represents the force generated when rotating an object (e.g., lower leg) about an axis (e.g., knee) (McGinnis, 2007). A strength index score is the sum of the scores generated for each joint. There is limited published research using isokinetic testing in an occupational setting. However, some research found a relationship between isokinetic test scores and injury reduction (Gilliam & Lund, 2000; Karwowski & Mital, 1986). Research comparing isokinetic tests with isometric and isotonic tests found the correlations among the tests to be high ($r = .91$ to .94) (Karwowski & Mital, 1986).

Muscular endurance tests assess the ability to withstand muscular fatigue. The duration of these tests varies in relation to the desired outcome and demands of the job. The arm endurance test, in which a subject pedals an arm ergometer at a set resistance level (e.g., 50 Watts) for a

specified time, is an example of a muscular endurance test (Gebhardt et al., 1998). The test is scored by counting the number of revolutions completed in a specified timeframe or assessing the duration for which a subject maintains a specific cadence. Other muscular endurance tests include sit-ups and push-ups.

Aerobic capacity tests assess the efficiency of the cardiovascular system to deliver oxygen to the muscles using a maximal or submaximal protocol. In a maximal test, the subject typically runs on a treadmill or pedals a bicycle at incremental workloads (e.g., increased speed and/or slope) until reaching exhaustion. The test uses a specific protocol (e.g., Bruce, Balke) and is scored as the time to exhaustion or an oxygen uptake value (i.e., $mL \bullet kg^{-1} \bullet min^{-1}$). The submaximal assessments include the step test, 1.5-mile run, 1-mile walk, 20-meter shuttle run, and bicycle test (e.g., YMCA, Astrand-Rhyming) (Astrand et al., 2003; Golding, 2000, Leger, Mercier, Gadoury, & Lambert, 1988; McArdle et al., 2015). The goal of submaximal tests is to provide an estimate of VO_{2max} using heart rate response to the exercise workload (e.g., step test), time to complete (e.g., 1.5-mile run), and/or distance covered (20-meter shuttle run). For tests involving heart rate response, the results are reported in $mL \bullet kg^{-1} \bullet min^{-1}$ and expressed as $VO_{2submax}$. For the timed and distance measures, tables are available for converting the measures to $VO_{2submax}$. Maximal and submaximal tests are used in employee selection. However, tests that measure physiological response (e.g., heart rate) are considered medical assessments by the Americans with Disabilities Act of 1990 (ADA) and the ADA Amendments Act of 2008 (ADAAA) and, therefore, should be given after a conditional job offer. Conversely, employers use aerobic tests that measure time or distance prior to a conditional job offer.

Flexibility and equilibrium tests, although used in employee selection, are rarely significant predictors of job performance. The correlations between job performance and these tests ranged from 0.00 to 0.18 (Baker & Gebhardt, 2012; Gebhardt et al., 2006). These correlations may reflect that flexibility does not contribute to successful performance of physical job tasks. Similarly, equilibrium tests were significantly related to job performance for only jobs requiring high levels of equilibrium (e.g., lashing containers to a ship at height of 40 feet) (Gebhardt, Baker, Volpe, & Younkins, 2010; Gebhardt, Schemmer, & Crump, 1985).

Some basic ability tests assess multiple abilities depending upon the intensity and duration of the test. For example, the arm endurance test described above can be a muscular endurance test or an anaerobic power test by shortening the duration (e.g., 10 seconds) and increasing the resistance (e.g., 100 Watts). Finally, basic ability tests are practical due to their small footprint, ease of storage, and transportability. The shortcoming of basic ability tests is that they do not resemble job tasks. Table 12.2 provides a listing of basic ability tests.

Job simulations or work sample tests include components of the job (e.g., pursuing a suspect, lifting boxes) and are used as predictors or criterion measures. Job simulations require performance of actual or simulated job tasks during the test. The primary advantage of job simulations is resemblance to the job. Further, they can be developed directly from the essential job tasks and provide an initial indication of how an individual handles equipment. The feasibility of developing a simulation that does not include equipment and skills learned in training or on the job may be difficult. However, substitution of non-job equipment (e.g., weight vest) for actual equipment (e.g., firefighter bunker gear) is possible. When job simulations consist of a series of tasks, the performance sequence, duration, and intensity should replicate the job as closely as possible. It is paramount that simulated tasks represent the critical physical job behaviors and working conditions and that the scoring metric is meaningful and identifies individual differences (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014). Many job simulations are time dependent, whereas others call for the completion of a task or set of tasks in a specific timeframe.

The primary disadvantages of job simulations are equipment size and construction, applicant safety (e.g., higher injury risk), and scoring metrics. Despite the disadvantages, job simulations based on essential tasks and ergonomic parameters relevant to the job possess content validity. Job simulations are more common for selection into public safety jobs than for other blue-color jobs (e.g., warehouse worker, pipefitter) and involve running or moving quickly, lifting, pushing, and pulling movements.

TABLE 12.2
Examples of Basic Physical Ability Tests

Physical Ability Muscular strength	*Example Tests*	*Physical Ability*	*Example Tests*
Upper body	Arm Lift	**Aerobic capacity**	Step test
	Shoulder Lift		1.5-Mile Run
	Handgrip		1-Mile Walk
	Static Push		20-meter shuttle run (Beep test)
	Static Pull		Bicycle ergometer (e.g., Astrand protocol) Treadmill (e.g., Bruce protocol)
	Chest Pull		
	Dynamic Lift		
	Push-Ups		
Trunk/core	Sit-Ups	**Flexibility**	Sit and Reach
	Trunk Pull		Joint range of motion
Lower body	Leg Lift Leg Press	**Equilibrium**	Stabiliometer Balance Beam
Muscular endurance			
Upper body	Arm Endurance	**Anaerobic power**	Shuttle Run
	Push-Ups		300-Meter Run
			Arm Ergometer (10 seconds)
			Margaria Test
			Illinois Agility Test
Trunk/core	Sit-ups		
Lower body	Stepping Platform Leg Endurance		

Another type of lifting test—isoinertial—assesses work capacity in a structured manner and increases the safety of the lifting tasks. Isoinertial tests encompass lifting predetermined weights from floor level to a defined height (e.g., waist, shoulder) at a specified pace (e.g., every 5 seconds). This differs from the psychophysical approach, in which the subject determines the weight lifted and the pace of lifting. Isoinertial tests increase the weight lifted by 5 or 10 pounds every 20 to 40 seconds. Depending upon the protocol, the weight lifted increases until the subject cannot complete the lift or the maximum weight defined by the job analysis is successfully lifted (Gebhardt et al., 2006; Mayer, Gatchel, & Mooney, 1990). Isoinertial tests are a reliable, safe, and inexpensive method to screen for jobs with frequent lifting (Hattori et al., 1998; Hazard, Reeves, & Fenwick, 1992; Lygren, Dragesund, Joensen, Ask, & Moe-Nilssen, 2005). Two studies found that the inclusion of an isoinertial lifting evaluation was more predictive of injuries than basic strength tests (Gebhardt et al., 2006; Mayer et al., 1990).

Factors to Consider in Test Development or Selection

When developing or selecting a physical performance test, one must consider the reliability, adverse impact, safety, and logistics related to test setup and administration. Myers et al. (1993) reviewed more than 20 basic ability tests and found the tests to be reliable with test-retest reliabilities ranging from 0.65 to 0.95. Reliability coefficients for job simulations tend to be similar to basic ability tests ($r = .50$ to .92). Research found lower reliabilities associated with lift/carry simulations ($r = .50$ to .57) and higher ones associated with task simulations such as manhole

hoist (0.83), ladder climb and carry (0.80–0.88), pursuit run (0.85–93), and pole climb (0.79) (e.g., Baker & Gebhardt, 2005; Gebhardt, Baker, & Volpe. 2012; Gebhardt et al., 1998).

Adverse impact by sex and age is a concern with physical tests. Due to physiological differences (e.g., lean body mass, percent body fat, height, weight), men perform significantly better on tests involving muscular strength, aerobic capacity, and anaerobic power, with effect sizes exceeding 1.0 (Blakley et al., 1994; Courtright, McCormick, Postlethwaite, Reeves, & Mount, 2013; Epstein et al., 2013; Gebhardt, 2007; Gebhardt & Baker, in press). Job simulations have greater sex differences than basic ability tests (Courtright et al., 2013; Gebhardt, 2007). With tests of flexibility and equilibrium, women performed similar or better than men (Gebhardt & Baker, 2010a). Studies that controlled for physiological differences (e.g., lean body mass) had mixed results, with some narrowing the gap and others showing significantly higher scores for men (Arvey, Landon, Nutting, & Maxwell, 1992; McArdle et al., 2015). However, this does not obviate the fact that women's mean performance on physical tests is significantly lower than men's performance. Since these tests are predictive of job performance, low test scores can lead to inadequate performance of physical job tasks, which can have severe consequences.

The physiological literature is replete with data showing decrements in physical performance with age (Baker & Gebhardt, 2012; Blakely et al., 1994; McArdle et al., 2015). These differences occurred for basic ability tests, job simulations, and job performance measures (Gebhardt & Baker, 2012).

In a large study of 50,000 men in blue-collar and public safety jobs, Baker (2007) found differences across ethnic groups. White men performed better than African American men on basic ability and job simulation tests requiring quick and/or continuous movement (e.g., pursuit run, 1.5-mile run, arm endurance, firefighter evolution), and White and African American men were significantly better than Hispanic men on strength tests (Baker, 2007; Blakely et al., 1994).

Although mean differences were present by sex, age, and ethnic group, examination of test performance using differential prediction found most physical tests fair across sex, ethnic group, and age subgroups (Baker & Gebhardt, 2012). While researchers readily recognize sex differences in physical test scores, reducing adverse impact on women, without compromising effective and safe job performance, is an issue. One approach is selecting and/or designing tests that have less adverse impact, after reviewing the validity, reliability, and adverse impact of current physical tests. Choosing a basic ability or job simulation test with less adverse impact is the first step. However, these choices may be limited if, for example, the job requires considerable upper body strength (e.g., lineworker). When designing new tests, the pilot and test samples must include an adequate number of women. Although more women perform nontraditional jobs, women make up less than 20% of the workers in physically demanding occupations (Department of Labor, 2015). Thus, organizations should recruit women, when feasible, to participate in validation studies. One validation study recruited women firefighters from neighboring states (Gebhardt & Baker, 1999). Another study recruited women soldiers to participate in validation research of military occupational specialties previously not open to women (Foulis et al., 2015). Without the women's data in these examples, identification of a fair and sound passing score would not have been possible.

The research literature shows tests of muscular endurance, flexibility, equilibrium, and coordination have lower sex differences than muscular strength with women performing similar to men on flexibility and equilibrium measures (Baker & Gebhardt, 2012; Courtright et al., 2013; Gebhardt, 2007; Gebhardt & Baker, 2010a, McArdle et al., 2015). Sex differences exist for aerobic capacity measures but are reduced by normalizing for body weight (i.e., $mL \cdot kg^{-1} \cdot min^{-1}$), which is appropriate when a known level of oxygen uptake is required to perform a task. Data reported on wildland firefighters and other jobs showed that a segment of women are capable of performing the same arduous tasks as men but have a greater energy expenditure. Gaskill et al. (2001) evaluated uphill hikes with firefighting equipment and found women and men completed the task successfully. However, women had a higher energy output and worked at almost 80% of VO_{2max}. Past research indicated that sustained work is typically performed at 40–50% of maximum (Astrand, et al., 2003; McArdle, et al., 2015). Thus, the women in the study could perform the task sequence one time, but sustaining the work over extended time periods may be difficult.

Past research found sex differences to be slightly less for basic ability tests versus job simulations, but when viewed as composite test batteries the sex effect sizes (*d*) were similar (Baker & Gebhardt, 2012; Courtright et al., 2013). Finally, research found sex differences present for not only physical tests but also measures of physical job performance (Courtright, et al., 2013; Gebhardt & Baker, 2010a; Hogan, 1991a). Thus, the differences were not due to test bias.

In addition to reliability and adverse impact issues, one must consider the safety and logistics associated with administering physical tests. Basic ability tests provide a controlled environment and are safer than job simulations. However, safe administration of job simulations is possible when organizations monitor test conditions such as floor surface, temperature, and the general testing environment.

VALIDITY OF PHYSICAL PERFORMANCE TESTS

Several approaches are available to establish the validity of physical tests. One involves establishing an empirical relationship between the test and a criterion. A second encompasses gathering of evidence that the test components have a verifiable link to job content or requirements by confirming the tests' relationship to a construct (e.g., muscular strength) or tasks required for job performance. Other methods include test transportability, job component validity, and synthetic validity. Baker and Gebhardt (2012) provide a description of these approaches and their use in physical assessments. The Uniform Guidelines (1978) and the Society of Industrial and Organizational Society (SIOP) *Principles* (SIOP, 2003) outline these validity approaches.

Numerous studies found physical performance tests to be valid predictors of job performance (Arvey et al., 1992; Blakely et al., 1994; Gebhardt, 2000; Hogan, 1991a). Prior research demonstrated the validity of basic ability and job simulation assessments (Arvey et al., 1992; Baker & Gebhardt, 2012; Blakely et al., 1994; Courtright, et al., 2013; Gebhardt, 2000; Gebhardt & Baker, 2010a, 2010b; Hogan, 1991a). Assessments of muscular strength, muscular endurance, aerobic, and anaerobic tests (e.g., arm lift, 1-mile run, sit-ups) had the highest relationship to job requirements for public safety and blue-collar positions, with flexibility and equilibrium occasionally contributing to the prediction of job performance. Our literature review found the simple validities for basic ability and job simulation tests range from 0.02 to 0.81 and 0.37 to 0.63, respectively, when using criterion measures such as supervisor/peer ratings and/or work simulations (e.g., Arvey et al., 1992; Gebhardt & Baker, 2010a; Hogan, 1991a).

When physiological and productivity measures were used to define job performance, the validities were comparable to other criterion measures. A study involving order fillers used productivity data (e.g., percent of the engineered standard) to identify a physical test battery (Gebhardt, Baker, Volpe, & Billerbeck, 2009). The significant simple validities ranged from 0.17 to 0.22 and increased to 0.29 to 0.38 when combined with supervisor ratings. Further, the correlation of the productivity measure with a work sample criterion measure was similar to the predictor tests in the final battery ($r = -0.24$). Other studies that used physiological measures (e.g., heart rate response, VO_2) found higher validities (e.g., 0.33 to 0.67) (Gebhardt et al., 2009; Sothmann et al., 2004).

When conducting criterion-related validity studies, creation/selection of the criterion measure(s) is as important as test selection. In addition to the criteria mentioned above, injury and lost workdays data are viable measures, but require large samples and may be confounded with organizational safety initiatives implemented concurrently with the testing. Regardless of the type of criteria used, the reliability of the measure should be determined (e.g., test-retest, Chronbach's alpha).

When using content validity, as in job simulations, the test must include tasks that replicate the job conditions, duration, and intensity of job tasks. The challenge when using a content model is establishing an accurate passing score. In light of the *Lanning v. SEPTA* (1999, 2002) litigation, there is added responsibility on an organization to gather empirical data (e.g., arrest rates) to establish a passing score that reflects the minimally acceptable level of job performance and is consistent with business necessity. If the evidence is insufficient to meet these criteria, the test will not withstand legal scrutiny (e.g., *EEOC v. Dial Corporation*, 2006).

In summary, determining the validity model to use depends on several factors: (a) type of test desired, (b) availability of job performance information (e.g., ratings, productivity, attrition), and (c) organizational resources. When criterion data are required (e.g., supervisor or peer ratings, productivity, attrition, injury data), the availability of personnel, probability of obtaining individual differences, type of data available (e.g., quantitative versus qualitative), and potential for confounding effects of other organizational programs must be considered. Generating usable workplace performance measures remains a challenge (Pulakos & O'Leary, 2011; Wigdor & Green, 1991). However, every effort should be made to identify valid and reliable workplace measures.

Selection of Final Test Battery

Selection of a final test battery requires knowledge of the job(s), the prediction space, and physical tests. If empirical data relating the predictor tests to a criterion measure are available, various statistical procedures exist to establish the test validity and test battery components. The first assessment should be a review of the correlations between the predictor tests and criterion measure(s) to identify potential tests. Depending upon the goal and constraints of the study, multiple statistical procedures (e.g., multiple regression, logistic regression, canonical correlation, regression tree) are available to identify a test battery. Multiple regression models used to identify tests that significantly add to the prediction of job performance help decrease the potential for test redundancy (e.g., highly correlated tests). This helps reduce use of two highly correlated tests (e.g., upper body strength test), which can increase adverse impact.

Other statistical procedures such as logistic regression allows for use of multiple predictors but requires a dichotomous criterion measure such as the level of aerobic capacity required to perform an order filler or firefighter job, or the likelihood of injury (Gebhardt et al., 2006; Hodgdon & Jackson, 2000; Pedhazur, 1997; Sothmann et al., 2004). Canonical correlation yields a correlation of two latent variables, one representing a set of independent variables, the other a set of dependent variables (Levine, 1977; Tabachnick & Fidell, 1997). This method allows the researcher to investigate a set of dependent variables instead of one variable. Each of these methods has advantages and disadvantages. Selection of a statistical technique is dependent on the available data, types of tests and criterion, organizational goals, and business necessity.

Physical performance tests commonly demonstrate adverse impact against women and older individuals in terms of test scores and passing rates. Therefore, it is important to establish test fairness across protected groups. A moderated regression analysis allows for examination of subgroup differences (Bartlett, Bobko, Mosier, & Hannan, 1978; Cleary, 1968). Research using this procedure found physical tests to be fair across sex, ethnic group, and age subgroups (Gebhardt et al., 1998; Sothmann et al., 2004).

TEST SCORING AND ADMINISTRATION

Types of Scoring

Two types of scoring methods commonly used for physical performance test batteries are multiple-hurdle (passing score for each test) and compensatory (sum of test scores) models. A third approach combines the compensatory and multiple-hurdle models and reduces the level of benefit for offsetting poor performance on one test with better scores on other tests found in the compensatory model.

Compensatory models, whether alone or in combination with a multiple hurdle model, normally result in less adverse impact against women than the multiple-hurdle approach (Baker & Gebhardt, 2012). When using a compensatory model, one must consider whether equal weighting (e.g., z-score) or multiple regression beta weights are most suited for the test battery.

When using a compensatory model, the simple raw score sum can be used if the beta weights from a regression equation are applied. Conversely, use of unit weighting requires a transformation of test scores due to the different scoring metrics of physical tests (e.g., seconds, pounds) and magnitude ranges of test scores. The third scoring model, which combines the multiple-hurdle and compensatory models, alleviates compensation for an extremely low score on one test by high scores on other tests, while maintaining the advantage of the compensatory model.

The third scoring model converts scores for each test in a battery into point values across a specific point range (e.g., stanine percentile). The sum of the points achieved across the tests produces a final test battery score. The point value ranges are identical for each test to allow for equal contribution of each test in the battery or are different and incorporate a weighting factor (e.g., beta weights). In this model, scores below specified levels receive zero points (Baker & Gebhardt, 2012). With this scoring model, a participant must meet or exceed the combined passing score and receive at least one point on each test. Two issues arise when attempting to use this method with multiple tests: (a) identification of the bandwidth for test scores assigned the same point value and (b) number of point values utilized per test. The bandwidth should be generated using data from test scores and take into account the statistical properties of the tests (e.g., standard error of the difference) (Cascio, Outtz, Zedeck, & Goldstein, 1991). Regardless of the scoring approach, the identification of a minimum acceptable level of performance and link to the passing score is critical.

Establishing Passing Scores

In the employment setting, passing scores identify individuals who are capable of performing or being trained to perform essential job tasks. Two basic types of passing scores, criterion-referenced and norm-referenced, are prevalent in physical testing (Landy & Conte, 2007; Safrit & Wood, 1989). The Uniform Guidelines (1978) indicated that passing scores should be "reasonable and consistent" with proficient job task performance. Criterion-referenced passing scores are best suited for meeting this goal. Use of expert judgment (e.g., Angoff) is one approach to identify passing scores, but data from concurrent and/or predictive validation studies help maximize test prediction.

Ergonomic and physiological data can provide actual values for completion of the work and in turn a passing score for a test. Sothmann and colleagues determined the minimum level of aerobic capacity required to perform firefighter tasks (e.g., pulling down ceiling) and used these data to establish the minimum aerobic capacity for a firefighter selection test (Sothmann et al., 1990; Sothmann et al., 2004). Similarly, direct measurement (e.g., force) of tasks involving muscular strength (e.g., tighten a turnbuckle) have been used to define successful and unsuccessful performance (Gebhardt et al., 1985; Gledhill & Jamnik, 1992; Jackson, Osburn, Laughery, & Vaubel, 1992). Absent these types of data, one must use a combination of expectancy and contingency tables, job analysis information, organizational preferences (e.g., test type), and business necessity to identify a passing score that maximizes prediction and minimizes adverse impact on protected groups.

In most instances, criterion-referenced passing scores are set using incumbent data. Cascio, Alexander, and Barrett (1988) stated that use of incumbents who are older and more experienced might lead to test score differences between incumbents and candidates. Research found older workers had lower physical test scores and performed at lower levels on physical job tasks than did younger workers (20–39 years), thus negating this concern (Baker & Gebhardt, 2012; Gebhardt et al., 1998).

For jobs that are time-sensitive (e.g., law enforcement, fire suppression, emergency medical service), the pace with which an individual responds is important to effective performance. For example, firefighters do not run when performing fire suppression activities, however, moving too slowly may result in lost lives and property. Experienced emergency personnel know the paces at which effective incumbents perform a job. Thus, pacing information provides an avenue for establishing a minimum requirement on time-sensitive job simulations. Several studies used pacing data to determine the passing scores for firefighter job simulations (Palmer, Baker,

Gebhardt, Abrams, & Weiner, 2014; Sothmann et al., 2004). In each study, researchers generated videotapes of a firefighter evolution completed at paces ranging from very fast to very slow based on incumbent performance. Samples of experienced firefighters viewed videotapes of varying performance levels and identified acceptable and unacceptable paces. This process resulted in the passing score corresponding to the slowest pace identified as meeting minimum job requirements.

Passing scores for physical ability tests in the public and private sectors are the same for all candidates regardless of age or sex, with the exception of selected law enforcement agencies that utilize normative data. For example, men age 20–29 years complete 40 sit-ups, whereas women age 20–29 years complete 35. Typically, the rationale for using normative sex and/or age data as passing scores is the premise that the agency is measuring physical fitness and not job performance. Recent mandates by Congress and the DoD resulted in single passing scores established for entry into military occupations previously not open to women, ensuring that both men and women possess the minimum physical qualifications. This decision and selection practices in non–law enforcement jobs clearly indicates the desire to ensure new hires are capable of meeting the physical demands of the jobs regardless of their age or sex.

Recent Federal District and Appeals Court decisions were mixed in terms of the legality of using different passing scores for subgroups (*Bauer v. Holder*, 2014; *Bauer v. Lynch*, 2016; *Easterling v. State of Connecticut Department of Correction*, 2011). In the Easterling case, female plaintiffs challenged the 1.5-mile run test stating that sex- and age-normed passing scores violated the Civil Rights Act of 1991. The court agreed with the plaintiffs that separate passing scores were not representative of minimum job requirements and stated:

> By definition, cutoff times that vary by gender and age cannot represent a measure of the minimum aerobic capacity necessary for successful performance as a CO. Only a single cutoff time could meet this standard.

In the Bauer case, the male plaintiff failed the FBI training academy test for the men's standard but would have passed using the women's standard (*Bauer v. Holder*, 2014). The District Court ruled that separate standards by sex were discriminatory and the judge stated:

> Female law enforcement officials perform the same physical job tasks as their male counterparts, gender-normed physical fitness standards cannot logically be used to measure an applicant's ability to perform discrete tasks such as restraining or chasing a suspect.

However, on appeal the court found the legal standard applied was incorrect, vacated the lower court decision, and remanded the case back to the district court (*Bauer v. Lynch*, 2016). The Easterling and initial Bauer decisions questioned how separate sex- and age-normed passing scores were relevant to meeting minimum job requirements. The defendants in the Bauer case argued that the different passing scores had no detrimental effect on men or women since the passing rates by sex were similar. The court in *Lanning v. SEPTA* (1999) considered the premise of passing scores linked to minimum job requirements and indicated that sex-normed scores could be pursued as long as the different passing scores could be linked to minimally acceptable job performance. Since there was no evidence that separate scores reflect minimally acceptable job performance, this proposal was not accepted. This is reminiscent of earlier litigation in which the court upheld use of norm-referenced tests on the basis that the tests were assessing fitness and not job requirements (*Alspaugh v. Michigan Law Enforcement Officers Training Council*, 2001; *Peanick v. Morris*, 1996). At this point, there are two conclusions related to use of normed passing scores. First, only selected law enforcement agencies use normed passing scores. Other public safety (e.g., fire and police departments) and private sector organizations use single passing scores. Baker's (2015) review of physical selection tests for state police agencies showed that the 79.6% of these agencies used basic ability tests and less than half had used sex-normed passing scores. Second, employers using single passing scores link passing scores to job performance requirements. This is consistent with the Uniform Guidelines (1978), which state that a passing score must represent minimally acceptable job performance. Proponents of normed passing scores do not address the need to represent minimally acceptable job performance.

Administration of Tests

Administration of physical tests, whether using sophisticated equipment or not, requires explicit test instructions, defined administrator and examinee procedures, and retest policies. Test instructions must provide adequate detail to ensure the examinee understands the purpose and goal of the test (e.g., complete maximum number of revolutions) and consequence of committing errors (e.g., repeat trial, fail test). Test administrator training programs must include procedures for testing examinees, use of test equipment, recognizing and demonstrating testing errors, and scoring the tests (e.g., time, count). When job simulations are used, administrators must practice cuing the examinee to the next test component, because improper timing of test cues impacts examinee performance.

Administrators and others should not provide encouragement to examinees because external motivation can alter performance. Testing examinees separately removes the possibility of external motivation and prevents subsequent examinees (e.g., second, third) from gaining test insight (e.g., pace) that was not available to the first examinee.

Placement of physical tests in the selection continuum and retest policies vary in relation to business necessity and type of test used. Organizations use physical tests either before or after a conditional job offer. If tests measure physiological parameters (e.g., heart rate) to generate a score, the ADA (1990) and ADAAA (2008) considers this a medical assessment, and the test must be given after a conditional job offer. Retest policies vary but should include information related to the minimum time needed to alter an individual's physiological state (e.g., muscular strength) and organizational needs and policies. From a physiological standpoint, 2–6 months of sustained exercise may be required to realize gains in strength and aerobic capacity to meet job requirements (McArdle et al., 2015; Nindl, 2015). Although retesting can effectively take place 3 months after initial testing, an organization may determine that the logistics for retesting are difficult or the pool of qualified applicants is sufficient. Conversely, the "shelf life" of test results might be affected by inactivity, injury, or aging for individuals who were not initially selected for the job. Therefore, a retest may be appropriate prior to entry into the job.

Physical Test Preparation

Physical training programs designed to increase job performance resulted in increases in muscular strength, muscular endurance, and aerobic capacity for women and men (Gebhardt & Crump, 1990; Jamnik, Thomas, & Gledhill, 2010; Knapik & Sharp, 1998; Roberts, 2009). Although the training programs increased women's muscular strength and aerobic capacity, the difference in performance between the sexes remained similar or became greater (Courtright et al., 2013). However, individuals who successfully completed these programs had a higher likelihood of meeting the minimum standards for a job than those who did not (Gebhardt & Baker, in press; Hogan & Quigley, 1994; Jamnik, Thomas, & Gledhill, 2010; Knapik et al., 2006). This is important for women seeking arduous jobs. Both general (e.g., weight lifting) and task-specific training programs produced increased fitness levels and the probability of meeting minimum job requirements (Jamnik, Thomas, Gledhill, 2010; Knapik & Sharp, 1998; Knapik et al., 2006). However, task-specific training provided better performance on job simulation tests. Thus, the type of training program used depends upon the physical test components and job tasks. More effective programs for women were staffed by trainers and included three to five sessions per week (Jamnik, Thomas, & Gledhill, 2010; Knapik et al., 2006). When job simulation tests are used, applicant practice sessions or instructional material (e.g., video, DVD) that outlines the test were effective preparation techniques (Hogan & Quigley, 1994; Sothmann et al., 2004).

Finally, one must remember that sex differences in physical performance persist even with training. For example, there is a 15–20% difference in aerobic capacity in trained athletes even when expressed relative to body weight ($mL \bullet kg^{-1} \bullet min^{-1}$). Elite women cross-country skiers have 15% less aerobic capacity than their male counterparts (McArdle et al., 2015). Jamnik and colleagues (2010) found similar results for a correctional officer applicant test preparation program in which men had a greater improvement in passing rate than women.

LEGAL ISSUES

Women being denied the opportunity to enter higher-paying trades and public safety jobs resulted greater scrutiny of selection procedures for arduous jobs. Litigation in the physical testing area focused mainly on adverse impact in the selection setting, with a few cases related to job retention. As stated above, the physiological sex differences led to test score differences, and in turn, disproportionate hiring of women. These test differences were not due to test bias, because there were corresponding differences for the criterion measure of interest (Hogan, 1991a). In fact, almost all physical tests violate the four-fifths rule, which defines adverse impact as the passing rate of a protected group (e.g., women) being less than 80% (four-fifths) of the majority group (e.g., men) (EEOC, 1978). Although almost all physical tests have an adverse impact on women, courts upheld the tests when the validity evidence demonstrated the relationship of the test and passing score(s) to the job (e.g., *Ernst v. City of Chicago*, 2015; *Porch v. Union Pacific Railroad*, 1997). However, when job analysis and/or validity evidence was lacking, the courts found for the plaintiff (e.g., *United States v. City of Erie*, 2005; *Varden v. City of Alabaster*, 2004). Prior papers have reviewed physical testing litigation (Hogan & Quigley, 1986; Terpstra, Mohamed, & Kethley, 1999). This review focuses on recent physical testing litigation in relation to several employment related laws. The Civil Rights Act (1964, 1991) and ADA (1990) are similar in requiring job-relatedness of selection procedures. In addition, ADA requires identification of a reasonable accommodation if available.

ADA OF 1990

Congress designed the ADA (1990) and amendments (2008) to protect individuals with disabilities in the private and nonfederal sectors. In the federal sector, the Rehabilitation Act of 1973 is a corollary to the ADA. Title I of ADA states that health/medical status (e.g., heart rate, blood pressure) inquires must follow conditional offer of employment, but that physical tests can be given prior to conditional job offer. These stipulations affect the type of assessments used for pre-job offer testing and the screening procedures used prior to test participation. For example, submaximal aerobic capacity tests (e.g., step, bicycle, treadmill) require monitoring heart rate and are not applicable in the pre-offer stage. Due to the inherent safety issues in physical testing, the ACSM recommends screening (e.g., blood pressure) prior to participation in exercise/testing (Pescatello et al., 2014). Because of the ADA medical test restrictions, employers use waiver forms and medical certification by a physician for pre-offer testing and medical examinations for the post-offer testing. It should be noted that a waiver does not absolve the employer of responsibility (*White v. Village of Homewood*, 1993).

Most ADA litigation dealt with medical issues (e.g., vision, diabetes, bipolar disorder) and incumbent personnel, rather than physical performance issues (Rothstein, Carver, Schroeder, & Shoben, 1999). The court cases involving incumbents showed that the employer must consider factors related to (a) involvement of health care personnel, equipment, or setting (EEOC, 2000; *Indergard v. Georgia-Pacific Corporation*, 2009) and (b) job requirements and physiological responses (*Andrews v. State of Ohio*, 1997; *Smith v. Des Moines*, 1996). In Indergard, the court determined what constitutes a physical test versus a medical examination and ruled in favor of the plaintiff. In Andrews and Smith, the court ruled that incumbent public safety employees who failed to meet physical standards were not disabled, just unfit for the job.

PHYSICAL TESTING LITIGATION

Litigation in the physical testing arena is affected by Title VII of the Civil Rights Act of 1964, Civil Rights Act of 1991 (CRA-91), the ADA (1990), and the Age Discrimination in Employment Act of 1967 (ADEA). Although Title VII set the initial standards for discrimination, the ADA had a profound effect on testing in the selection setting (Gutman,

Koppes, & Vodanovich, 2011). Review of physical testing litigation showed four premises governed whether the court upheld or struck down a test. These were the (1) plaintiff's ability to show the test had adverse impact on a protected group, (2) defendant's ability to show the test was job related, (3) defendant's need for business necessity, and (4) plaintiff's ability to show the existence of an alternative assessment with less adverse impact and equal validity (Gutman et al., 2011). In addition to these premises, the courts found the quality of some studies did not meet the Uniform Guidelines (1978) parameters and/or job relatedness burden of proof (e.g., *United States v. City of Erie*, 2005).

Reviews of physical testing litigation found that these tests were struck down more often than upheld in the 1970s and 1980s (Baker & Gebhardt, 2012; Hogan & Quigley, 1986) due to lack of or faulty job analyses or low quality of validation studies. During this period, the courts accepted content validity of a job simulation test based on detailed job analysis (*Hardy v. Stumpf*, 1978) but did not for basic ability tests (*Berkman v. City of New York*, 1982).

After the 1980s, employers were more successful in defending their physical tests (Baker & Gebhardt, 2012). This was attributed to the enactment of the EEOC Uniform Guidelines (1978), which provided guidance for conduct of job analysis and validity studies. However, when a defendant failed to meet these criteria, the plaintiffs prevailed with the court, citing problems with the job analysis, test and validity, and business necessity (*Legault v. Russo*, 1994; *United States v. City of Eri*e, 2005).

More recently, the courts ruled on issues related to physical test development, use of passing scores, and business necessity. The *Standards for Educational and Psychological Testing* (AERA et al., 2014) and SIOP *Principles* (2003) address providing empirical data that identify the relationship of the test to relevant criteria and minimum job requirements. In *Ernst et al. v. City of Chicago* (2015), the women plaintiffs charged disparate impact and treatment after failing a paramedic basic ability test. The court found that the test was job related and consistent with business necessity. The plaintiffs failed to demonstrate that the employment practice caused disparate impact on the basis of sex and were unable to provide equally valid alternatives. Finally, the jury found no evidence of disparate treatment.

The *Lanning v. SEPTA* (1999, 2002) cases and their impact on passing scores received a great deal of attention (Gutman, 2003; Sharf, 1999, 2003). This litigation centered around adverse impact on women who failed to complete the 1.5-mile run in 12 minutes. To improve the law enforcement capabilities of their police force, SEPTA implemented a physical performance test battery that included a 1.5-mile run. After the appeals court remanded the initial decision, the lower court found for the defendant and determined that the passing score reflected minimally acceptable performance defined as "likely to do the job," not "some chance of doing the job" (Sharf, 2003). Although the Lanning case showed that various information sources are acceptable for defending a test and passing score, it suggested that the Uniform Guidelines and SIOP *Principles* (2003) were not necessarily relevant to establishing job-relatedness. The Lanning ruling applied a stricter burden to prove job-relatedness and business necessity of a test.

In cases where the plaintiff prevailed, issues related to business necessity and job-relatedness were at the forefront, or the court accepted the plaintiff's less discriminatory alternative. The court determined that a job simulation involving lifting bars to selected heights was more difficult than the job and discriminated against women based on the passing score and/or subjective judgment of their performance (*EEOC v. Dial Corp*, 2006). The court denied Dial's business necessity defense of injury reduction because they could not determine whether injury reduction was a result of the test or other organizational interventions (e.g., safe lifting). When the plaintiff prevails, the recurring theme centers on job analysis and job-relatedness regardless of the type of test used (e.g., *United States v. City of Erie*, 2005). In one case, plaintiffs identified a less discriminatory alternative test and settled out of court with the city (*Vasich v. City of Chicago*, 2013).

Other challenges to physical testing relate to employee retention or promotion in fire and law enforcement departments. The courts ruled that an employer can institute incumbent physical assessments, but these assessments must stand up to legal scrutiny in regard to validity and job relatedness (*Fraternal Order of Police v. Butler County Sheriff Department*, 2006; *Pentagon Force Protection*

Agency v. Fraternal Order of Police, 2004; *Smith v. Des Moines*, 1997; *Varden v. City of Alabaster*, 2004). In the private sector, an arbitrator upheld the use of physical tests for incumbent job transfers to physically demanding jobs (*UWUA Local 223 v. The Detroit Edison Co.*, 1991). In addition to disparate impact by sex, incumbent testing also addresses age and disability discrimination (ADA, 1990; ADEA, 1967). In two state police cases, incumbent personnel brought suit under ADEA against the Commonwealth of Massachusetts (*Gately v. Massachusetts*, 1992, 1996) and the State of Vermont (*Badgley v. Walton*, 2010). In both cases, the courts upheld the states' procedures that utilized physical tests to assess incumbents.

Finally, as outlined above under scoring procedures, the courts have provided mixed decisions regarding sex- and/or age-normed physical test passing scores (*Bauer v. Holder*, 2014; *Bauer v. Lynch*, 2016; *Easterling v. State of Connecticut Department of Correction*, 2011). We will stay tuned for decisions on age and sex norming. In summary, job analysis, job-relatedness, and business necessity were the primary issues in the court decisions. In reviewing case law, we found that the defendant prevailed 60–80% of the time depending upon the type of test used (e.g., basic ability, job simulation) and the type of job.

BENEFITS AND TRENDS IN PHYSICAL TESTING

The benefits of physical testing for selection into arduous jobs range from reduction in lost work time, turnover, and injuries to increases in productivity. Studies have demonstrated the relationship between physical capabilities and injuries and productivity (Knapik et al., 2011, Sackett & Mavor, 2006). In a longitudinal study, the military demonstrated reduction in injuries in basic training by using physical testing to identify individuals who were unable to meet the training demands (Knapik et al., 2007). Knapik et al. (2011) showed significantly fewer injuries sustained in defensive tactics training and other physical tasks for individuals in the top three quartiles of muscular strength and aerobic capacity in a law enforcement academy. Women and men in the lowest quartile for muscular strength and aerobic capacity were 1.51 to 1.53 times and 1.39 to 2.01 times, respectively, more likely to be injured. Similar injury reductions were found for tree planter, wildland firefighter, and manual materials handling incumbents with higher strength and aerobic capacities (Craig, Congleton, Kerk, Amendola, & Gaines, 2006; Gilliam & Lund, 2000; Roberts, 2009; Sharkey & Gaskill, 2009).

When using pre-employment physical tests, researchers found injuries and days lost from work decreased. One study examined 5 years of injury and time loss data in the railroad industry using tested and hired (n = 12,714) and not tested and hired (n = 15,794) train service samples (Baker & Gebhardt, 2001). The tested group had fewer injuries than the not tested group (648 vs. 3,898). When age and tenure were controlled, the results showed significant differences for days lost (tested = 77.2; not tested = 142.4) and cost per injury (tested = $15,315; not tested = $66,148). Research in the freight industry found significantly fewer lost workdays for the tested group than the not-tested group (Baker, Gebhardt, & Koeneke, 2001). In the warehouse industry, individuals who passed a physical selection test met production standards faster than non-tested new hires and had lower turnover rates (S. Bolin, personal communication, November 20, 2015).

Several trends evolved in physical testing in the past few years. First, there are more data related to women's performance on physical tests and arduous job performance. Second, some organizations provide greater information to applicants in terms of test protocols and scoring metrics (Baker, 2015). For example, 100% of the state police listed the physical test requirements online, but only 23% of private sector organizations listed theirs (Baker, 2015; Baker, St. Ville, Gebhardt, & Volpe, 2014). Third, basic ability tests, job simulations, and combined ability-simulation batteries remain viable physical test formats. This finding reflects the number of state police agencies with basic ability tests (68%), job simulations (18%), and combination ability-simulation tests (14%) (Baker, 2015). Similar results were found for a review of physical tests across private and public sectors, with basic ability tests (55.8%) being most prevalent followed by combination tests (23.1%) and job simulations (21.1%) (Baker et al., 2014). Public safety and warehouse/distribution organizations accounted for the highest percentage using physical tests in the selection setting.

Fourth, more public safety agencies are assessing incumbent personnel. In some instances, this is due to nationwide policies such as the National Fire Protection Agency (NFPA) 1583 document (2015) that states the importance of annual assessment of firefighters' physical capabilities. In other instances, agencies strive to engender healthy lifestyles and reduce injuries. Fifth, physical test litigation is not declining and is based primarily on test score and selection ratio differences by sex.

In summary, physical tests developed and validated in accordance with the laws and professional standards benefit the employer and the employee by identifying individuals who are capable of meeting the physical demands of arduous jobs. Individuals who pass such tests are more likely to be successful performing physical work and less likely to incur worker compensation costs (e.g., lost workdays, injury). Physical tests, as with all tests, are subject to legal scrutiny and withstand legal challenge when a detailed job analysis is present, tests are job-related, and business necessity is met. In regard to physical training programs, both women and men benefit from these programs in relation to meeting minimum selection requirements. As more women enter nontraditional jobs, the knowledge base of women's performance on physical tests and performance outcomes in arduous jobs will increase.

REFERENCES

ADA Amendments Act of 2008 (Public Law 110–325, ADAAA).

Age Discrimination in Employment Act of 1967, 29 U.S.C. § 621 et seq. (1967).

Alspaugh v. Michigan Law Enforcement Officers Training Council, 634 N.W.2d 161 (Mich. App. 2001).

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Americans with Disabilities Act of 1990, 42 U.S. C. A.

Andrews v. State of Ohio, 104 F.3d 803 (6th Cir., 1997).

Arvey, R. D., Landon, T. E., Nutting, S. M., & Maxwell, S. E. (1992). Development of physical ability tests for police officers: A construct validation approach. *Journal of Applied Psychology, 77*, 996–1009.

Astrand, P., Rodahl, K., Dahl, H. A., & Stromme, S. G. (2003). *Textbook of work physiology* (4th ed.). Champaign, IL: Human Kinetics.

Badgley and Whitney v. Walton and Sleeper, Commissioners of Public Safety and Department of Public Safety, VT Supreme Court #2008–385, 2010.

Baker, T. A. (2007). *Physical performance test results across ethnic groups: Does the type of test have an impact?* Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology. New York, NY.

Baker, T. A. (2015). *Review of law enforcement physical testing at the state level.* Alexandria, VA: HumRRO.

Baker, T. A., & Gebhardt, D. L. (2001). *Utility of physical performance tests in reduction of days lost and injuries in railroad train service positions.* Beltsville, MD: Human Performance Systems.

Baker, T. A., & Gebhardt, D. L. (2005). *Examination of revised passing scores for state police physical performance selection tests.* Beltsville, MD: Human Performance Systems.

Baker, T. A., & Gebhardt, D. L. (2012). Chapter 13: The assessment of physical capabilities in the workplace. In N. Schmitt (Ed.), *Handbook of assessment and selection* (pp. 274–296). New York, NY: Oxford University Press, Inc.

Baker, T. A, Gebhardt, D. L., & Koeneke, K. (2001). *Injury and physical performance tests score analysis of Yellow Freight System dockworker, driver, hostler, and mechanic positions.* Beltsville, MD: Human Performance Systems, Inc.

Baker, T. A., St. Ville, K. A., Gebhardt, D. L., & Volpe, E. K. (2014). *Use of physical performance tests for selection in the private and public sectors* (White paper). Beltsville, MD: Human Performance Systems, Inc.

Bartlett, C. J., Bobko, P., Mosier, S. B., & Hannan, R. (1978). Testing for fairness with a moderated multiple regression strategy: An alternative to differential analysis. *Personnel Psychology, 31*, 233–241.

Bauer v. Holder, No. 1:2013cv00093—Document 125 (E.D. Va. 2014).

Bauer v. Lynch, Case No. 14–2323 (4th Cir. Jan 11, 2016).

Berkman v. City of New York, 536 F. Supp. 177, 30 Empl. Prac. Dec. (CCH) 33320 (E.D.N.Y. 1982).

Bilzon, J. L., Scarpello, E. G., Smith, C. V., Ravenhill, N. A., & Rayson, M. P. (2001). Characterization of the metabolic demands of simulated shipboard Royal Navy fire-fighting tasks. *Ergonomics, 44*, 766–780.

Blakley, B. R., Quinones, M. A., Crawford, M. S., & Jago, I. A. (1994). The validity of isometric strength tests. *Personnel Psychology, 47*, 247–274.

Bureau of Labor Statistics. (2011). *Employed persons by detailed occupation, sex, race, and Hispanic or Latino ethnicity*. Retrieved from http:// www.bls.gov/cps/cpsaat11.pdf

Carter, A. *All combat roles now open to women, Defense Secretary says*. Retrieved from http://www.nytimes.com/2015/12/04/us/politics/combat-military-women-ash-carter.html

Cascio, W. F., Alexander, R. A., & Barrett, G. V. (1988). Setting cutoff scores: Legal, psychometric, and professional issues and guidelines. *Personnel Psychology, 41*, 1–24.

Cascio, W. F., Outtz, J. L., Zedeck, S., & Goldstein, I. L. (1991). Statistical implications of six methods of test score use in personnel selection. *Human Performance, 4*, 233–264.

Chaffin, D. B., Herrin, G. D., Keyserling, W. M., & Foulke, J. A. (1977). *Pre-employment strength testing in selecting workers for materials handling jobs* (Report CDC-99–74–62). Cincinnati, OH: National Institute for Occupational Safety and Health, Physiology, and Ergonomics Branch.

Civil Rights Act of 1964 (Title VII), 42 U.S.C. §2000e-2 et seq., (1964).

Civil Rights Act of 1991, S. 1745, 102nd Congress (1991).

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement, 5*, 115–124.

Courtright, S. H., McCormick, B. W., Postlethwaite, B. E., Reeves, C. J., & Mount, M. K. (2013). A meta-analysis of sex differences in physical ability: Revised estimates and strategies for reducing differences in selection contexts. *Journal of Applied Psychology, 98*(4), 623–641.

Craig, B. N., Congleton, J. J., Kerk, C. J., Amendola, A. A., & Gaines, W. G. (2006). Personal and non-occupational risk factors and occupational injury/illness. *American Journal of Industrial Medicine, 49*, 249–260.

Davis, P. O., Dotson, C. O., & Santa Maria, D. L. (1982). Relationship between simulated fire fighting tasks and physical performance measures. *Medicine and Science in Sports and Exercise, 14*, 65–71.

Department of Labor. Quick Facts on Nontraditional Occupations for Women. (December 30, 2015). Retrieved from http://www.dol.gov/wb/factsheets/nontra2008.htm

Easterling v. State of Connecticut, Department of Correction, 783 F. Supp. 2d 323 (2nd Cir. 2011).

Epstein, Y., Yanovich, R., Moran, D. S., & Heled, Y. (2013). Physiological employment standards IV: Integration of women in combat units physiological and medical considerations. *European Journal of Applied Physiology and Occupational Physiology*, *113*, 2673–2690.

Equal Employment Opportunity Commission. (2000). *Enforcement guidance: Disability-related inquiries and medical examinations of employees under the Americans with Disabilities Act (ADA)*. Washington, DC: http://www.eeoc.gov/policy/docs/guidance-inquiries.html.

Equal Employment Opportunity Commission v. Dial Corp, No. 05–4183/4311 (8th Cir. 2006).

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice. (1978). *Uniform guidelines on employee selection procedures*. Washington, DC: Bureau of National Affairs, Inc.

Ernst et al. v. City of Chicago, No. 1:08-cv-4370 (N.D. Ill.. 2015).

Fleishman, E. A. (1964). *Structure and measurement of physical fitness*. Englewood, NJ: Prentice Hall.

Fleishman, E. A., Gebhardt, D. L., & Hogan, J. C. (1986). The perception of physical effort in job tasks. In G. Borg & D. Ottoson (Eds.), *The perception of exertion in physical work* (pp. 225–242). Stockholm, Sweden: Macmillan Press.

Fleishman, E. A., & Quaintance, M. K. (1984). *Taxonomies of human performance*. New York, NY: Academic Press.

Foulis, S. A., Redmond, J. E., Warr, B. J., Zambraski, E. J. Frykman, P. N., Gebhardt, D. L., Baker, T. A., & Sharp, M. A. (2015). *Development of a physical employment testing battery for 12B Combat Engineers*. Natick, MA: U.S. Army Research Institute of Environmental Medicine-Military Performance Division.

Fraternal Order of Police Local 101 v. Butler County Sheriff's Department, #05-ULP-09–0509, 23 OPER 30 (Ohio SERB, 2006).

Gaskill, S. E., Ruby, B. C., Walker, A. J., Sanchez, O. A., Serfass, R. C., & Leon, A. S. (2001). Validity and reliability of combining three methods to determine ventilatory threshold. *Medicine & Science in Sports & Exercise, 33*, 1841–1848.

Gately v. Massachusetts, 92-CV-13018-MA (D. Mass. Dec. 30, 1992).

Gately v. Massachusetts, No. 92–13018 (D. Mass. Sept. 26, 1996).

Gebhardt, D. L. (1984). *Revision of physical ability scales*. Bethesda, MD: Advanced Research Resources Organization.

Gebhardt, D. L. (2000). Establishing performance standards. In S. Constable & B. Palmer (Eds.), *The process of physical standards development*. Wright-Patterson Air Force Base, OH: Human Systems Information Analysis Center.

Gebhardt, D. L. (April 2007). *Physical performance testing: What is the true impact?* Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology. New York, NY.

Gebhardt, D. L., & Baker, T. A. (1999). *Validation of physical performance tests for the selection of firefighters in the State of New Jersey*. Beltsville, MD: Human Performance Systems.
Gebhardt, D. L., & Baker, T. A. (2006). *Determination of incumbent passing scores for the Massachusetts State Police physical performance test*. Beltsville, MD: Human Performance Systems.
Gebhardt, D. L., & Baker, T. A. (2010a). Physical performance. In J. C. Scott & D. H. Reynolds (Eds.), *Handbook of workplace assessment* (pp. 179–196). San Francisco, CA: Jossey-Bass.
Gebhardt, D. L., & Baker, T. A. (2010b). Physical performance tests. In J. Farr & N. Tippins (Eds.), *Handbook on employee selection* (pp. 277–298). New York, NY: Routledge.
Gebhardt, D. L., & Baker, T. A. (2012). *Examination of the effects of age on performance of physical requirements*. Beltsville, MD: Human Performance Systems.
Gebhardt, D. L., & Baker, T. A. (In press). Physical performance assessment. In S. G. Rogelberg (Ed.), *Encyclopedia of industrial and organizational psychology*. Thousand Oaks, CA: Sage.
Gebhardt, D. L., Baker, T. A., & Sheppard, V. A. (1998). *Development and validation of physical performance tests for BellSouth physically demanding jobs*. Hyattsville, MD: Human Performance Systems.
Gebhardt, D. L., Baker, T. A., & Thune, A. (2006). *Development and validation of physical performance, cognitive, and personality assessments for selectors and delivery drivers*. Beltsville, MD: Human Performance Systems.
Gebhardt, D. L., Baker, T. A., & Volpe, E. K. (2012). *Development and validation of physical performance tests for United States secret service special agents and uniformed division officers*. Beltsville, MD: Human Performance Systems, Inc.
Gebhardt, D. L., Baker, T. A., Volpe, E. K., & Billerbeck, K. T. (2009). *Development and validation of physical performance tests for selection of orderfillers*. Beltsville, MD: Human Performance Systems, Inc.
Gebhardt, D. L., Baker, T. A., Volpe, E. K., & Younkins, D. H. (2010). *Development and validation of physical performance tests for CSX Transportation physically demanding jobs. Volume 2: Test development and validation report*. Beltsville, MD: Human Performance Systems, Inc.
Gebhardt, D. L., & Crump, C. E. (1984). *Validation of physical performance selection tests for paramedics*. Bethesda, MD: Advanced Research Resources Organization.
Gebhardt, D. L., & Crump, C. E. (1990). Employee fitness and wellness programs in the workplace. *American Psychologist, 45*, 262–272.
Gebhardt, D. L., Schemmer, F. M., & Crump, C. E. (1985). *Development and validation of selection tests for longshoremen and marine clerks*. Bethesda, MD: Advanced Research Resources Organization.
Gilliam, T., & Lund, S. J. (2000). Injury reduction in truck driver/dock workers through physical capability new hire screening. *Medicine and Science in Sports and Exercise, 32*, S126.
Gledhill, N., & Jamnik, V. K. (1992). Characterization of the physical demands of firefighting. *Canadian Journal of Sport Science, 17*, 207–213.
Golding, L. A. (2000). *YMCA fitness testing and assessment manual* (4th ed.). Champaign, IL: Human Kinetics.
Guion, R. M. (1998). *Assessment, measurement and prediction for personnel decisions*. Mahwah, NJ: Lawrence Erlbaum.
Gutman, A. (2003). Adverse impact: Why is it so difficult to understand? *The Industrial-Organizational Psychologist, 40*, 50.
Gutman, A., Koppes, L. L., & Vodanovich, S. J. (2011). *EEO law and personnel practices* (3rd ed.). Thousand Oaks, CA: Sage Publications.
Hardy v. Stumpf, 17 Fair Empl. Prac. Cas. (BNA) 468 (Sup. Ct. Cal. 1978).
Hattori, Y., Ono, Y., Shimaoka, M., Hiruta, S., Kamijima, M., & Takeuchi, Y. (1998). Test-retest reliability of isometric and isoinertial testing in symmetric and asymmetric lifting. *Ergonomics, 41*, 1050–1059.
Hazard, R. G., Reeves, V., & Fenwick, J. W. (1992). Lifting capacity: Indices of subject effort. *Spine, 17*, 1065–1070.
Hodgdon, J. A., & Jackson, A. S. (2000). Physical test validation for job selection. In S. Constable & B. Palmer (Eds.), *The process of physical fitness standards development* (pp. 139–177). Wright-Patterson Air Force Base, OH: Human Systems Information Analysis Center.
Hogan, J. C. (1991a). Physical abilities. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 2, pp. 753–831). Palo Alto, CA: Consulting Psychologist Press.
Hogan, J. C. (1991b). Structure of physical performance in occupational tasks. *Journal of Applied Psychology, 76*, 495–507.
Hogan, J. C., & Quigley, A. M. (1986). Physical standards for employment and the courts. *American Psychologist, 41*, 1193–1217.
Hogan, J. C., & Quigley, A. M. (1994). Effects of preparing for physical ability tests. *Public Personnel Management, 23*, 85–104.
Indergard v. Georgia-Pacific Corporation, 582 F.3d 1049 (9th Cir. 2009).
Jackson, A. S., Osburn, H. G., Laughery, K. R., & Vaubel, K. P. (1992). Validity of isometric tests for predicting the capacity to crack open and closed industrial valves. In *Human factors and ergonomics society annual meeting proceedings* (pp. 688–691). Santa Monica, CA: Human Factors and Ergonomics Society.

Jackson, A. S., & Sekula, B. K. (1999). The influence of strength and gender on defining psychophysical lifting capacity. *Proceeding of the Human Factors and Ergonomics Society, 43*, 723–727.

Jamnik, V. K., Thomas, S. G., Burr, J. F., & Gledhill, N. (2010). Construction, validation, and derivation of performance standards for a fitness test for correctional officer applicants. *Applied Physiology, Nutrition, and Metabolism, 35*, 59–70.

Jamnik, V. K., Thomas, S. G., & Gledhill, N. (2010). Applying the Meiorin decision requirements to the fitness test for correctional officer applicants: Examining adverse impact and accommodation. *Applied Physiology, Nutrition, and Metabolism, 35*, 71–81.

Karwowski, W., & Mital, A. (1986). Isometric and isokinetic testing of lifting strength of males in teamwork. *Ergonomics, 29*, 869–878.

Knapik, J. J., Darakjy, S., Hauret, K. G., Canada, S., Scott, S., Rieger, W., Marin, R., & Jones, B. H. (2006). Increasing the physical fitness of low-fit recruits before basic combat training: An evaluation of fitness, injuries, and training outcomes. *Military Medicine, 171*, 45–54.

Knapik, J. J., Jones, S. B., Darakjy, S., Hauret, K. G., Bullock, S. H., Sharp, M. A., & Jones, B. H. (2007). Injury rates and injury risk factors among U.S. Army wheel vehicle mechanics. *Military Medicine, 172*, 988–996.

Knapik, J. J., & Sharp, M. A. (1998). Task-specific and generalized physical training for improving manual-material handling capability. *International Journal of Industrial Ergonomics*, *22*, 149–160.

Knapik, J. J., Spiess, A. S., Swedler, D., Grier, T., Hauret, K. G., Yoder, J., & Jones, B. H. (2011). Retrospective examination of injuries and physical fitness during Federal Bureau of Investigation new agent training. *Journal of Occupational Medicine and Toxicology*, *6*, 26–37.

Landy, F., Bland, R., Buskirk, E., Daly, R. E., Debusk, R. F., Donovan, E. et al. (1992). *Alternatives to chronological age in determining standards of suitability for public safety jobs* (Technical Report) University City, PA: Center for Applied Behavioral Sciences, Pennsylvania State University.

Landy, F. J., & Conte, J. M. (2007). *Work in the 21st century: An introduction to industrial and organizational psychology*. Malden, MA: Blackwell.

Lanning v. Southeastern Pennsylvania Transportation Authority, 181 F.3d 478, 482–484 (3rd Cir. 1999).

Lanning v. Southeastern Pennsylvania Transportation Authority, 308 F.3d 286 (3rd Cir. 2002).

Legault v. Russo, 64 FEP Cases (BNA) 170 (D.N.H., 1994).

Leger, L. A., Mercier, D., Gadoury, C., & Lambert, J. (1988). The multistage 20 metre shuttle run test for aerobic fitness. *Journal of Sports Sciences, 6*(2), 93–101.

Levine, M. S. (1977). *Canonical correlation analysis: Uses and interpretation*. Beverly Hills, CA: Sage.

Lygren, H., Dragesund, T., Joensen, J., Ask, T., & Moe-Nilssen, R. (2005). Test-retest reliability of the Progressive Isoinertial Lifting Evaluation (PILE). *Spine, 30*, 1070–1074.

Mayer, T., Gatchel, R., & Mooney, V. (1990). Safety of the dynamic progressive isoinertial lifting evaluation (PILE) test. *Spine, 15*, 985–986.

McArdle, W. D., Katch, F. I., & Katch, V. L. (2015). *Exercise physiology: Energy, nutrition, and human performance physiology* (8th ed.). Baltimore, MD: Wolters Kluwer Health | Lippincott Williams & Wilkins.

McGinnis, P. M. (2007). *Biomechanics of sport and exercise* (2nd ed.). Champaign, IL: Human Kinetics.

Myers, D. C., Gebhardt, D. L., Crump, C. E., & Fleishman, E. A. (1993). The dimensions of human physical performance: Factor analyses of strength, stamina, flexibility, and body composition measures. *Human Performance, 6*, 309–344.

National Fire Protection Agency (NFPA). (2015). *NFPA 1583: Standard on health related fitness programs for fire department members*. Quincy, MA: National Fire Protection Agency.

Nindl, B. C. (2015). Physical training strategies for military women's performance optimization in combat-centric occupations. *Journal of Strength Conditioning, 29*, S101–S106.

Palmer, P., Baker, T., Gebhardt, D., Abrams, J., & Weiner, J. (2014). *Validation study of the FDNY Academy Functional Skills Training and Testing (FST) and Practical Skills Test (PST)*. Burbank, CA: PSI.

Pandolf, K. B., Burse, R. L., & Goldman, R. F. (1977). Role of physical fitness in heat acclimatization, decay and reinduction. *Ergonomics, 20*, 399–408.

Peanick v. Morris (US Marshals Service), 95–2594 (8th Cir. 1996).

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). New York, NY: Harcourt Brace College Publishers.

Pentagon Force Protection Agency v. Fraternal Order of Police DPS Labor Committee, FLRA Case #WA-CA-04–0251 (Wash. Region, 2004).

Pescatello, L. S., Arena, R., Riebe, D., & Thompson, P. D. (2014). *ACSM's guidelines for exercise testing and prescription* (9th ed.). Philadelphia, PA: Lippincott Williams & Wilkins.

Porch v. Union Pacific Railroad, Administrative law proceeding, State of Utah 1997.

Pulakos, E. D., & O'Leary, R. S. (2011). Why is performance management broken? *Industrial and Organizational Psychology, 4*(2), 146–164.

Rehabilitation Act of 1973, 29 U.S.C. 701 et seq. (1973).
Reilly, T. J., Gebhardt, D. L., Billing, D. C., Greeves, J. P., & Sharp, M. A. (2015). Development and implementation of evidence-based physical employment standards: Key challenges in the military context. *The Journal of Strength and Conditioning Research, 29*(Suppl. 11), S28–S33.
Roberts, D. (2009). The occupational athlete: Injury reduction and productivity enhancement in reforestation workers. In N. P. Pronk (Ed.), *ACSM's worksite health handbook: A guide to building healthy companies*. Champaign, IL: Human Kinetics.
Rothstein, M. A., Carver, C. B., Schroeder, E. P., & Shoben, E. W. (1999). *Employment law* (2nd ed.). St. Paul, MN: West Group.
Sackett, P. R., & Mavor, A. S. (2006). *Assessing fitness for military enlistment: Physical, medical and mental health standards*. Washington, DC: The National Academies Press.
Safrit, M. J., & Wood, T. M. (1989). *Measurement concepts in physical education and exercise science*. Champaign, IL: Human Kinetics.
Sharf, J. C. (1999). Third Circuit's *Lanning v. SEPTA* decision: Business necessity requires setting minimum standards. *The Industrial-Organizational Psychologist, 37*, 149.
Sharf, J. C. (2003). *Lanning* revisited: The Third Circuit again rejects relative merit. *The Industrial-Organizational Psychologist, 40*, 40.
Sharkey, B. J., & Gaskill, S. E. (2009). *Fitness and work capacity*. Boise, ID: National Wildlife Coordinating Group.
Smith v. Des Moines, #95–3802, 99 F.3d 1466, 1996 U.S. App. Lexis 29340, 72 FEP Cases (BNA) 628, 6 AD Cases (BNA) 14 (8th Cir. 1996). [1997 FP 11]
Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.
Solianik, R., Skurvydas, A., Mickevičienė, D., & Brazaitis, M. (2014). Intermittent whole-body cold immersion induces similar thermal stress but different motor and cognitive responses between males and females. *Cryobiology, 69*, 323–332.
Sothmann, M. S., Gebhardt, D. L., Baker, T. A., Kastello, G. M., & Sheppard, V. A. (2004). Performance requirements of physically strenuous occupations: Validating minimum standards for muscular strength and endurance. *Ergonomics, 47*, 864–875.
Sothmann, M. S., Saupe, K., Jasenof, D., Blaney, J., Donahue-Fuhrman, S., Woulfe, T., et al. (1990). Advancing age and the cardiovascular stress of fire suppression: Determining the minimum standard for aerobic fitness. *Human Performance, 3*, 217–236.
Tabachnick, B. G., & Fidell, L. S. (1997). *Using multivariate statistics*. New York, NY: HarperCollins College Publishers.
Terpstra, D. A., Mohamed, A. A., & Kethley, R. B. (1999). An analysis of federal court cases involving nine selection devices. *International Journal of Selection and Assessment, 7*, 26–34.
Tikuisis, P., Jacobs, I., Moroz, D., Vallerand, A., & Martineau, L. (2000). Comparison of thermoregulatory responses between men and women immersed in cold water. *Journal of Applied Physiology*, *89*, 1403–1411.
Tipton, M. J., Milligan, G. S., & Reilly, T. J. (2013). Physiological employment standards I. Occupational fitness standards: Objectively subjective? *European Journal of Applied Physiology, 113*(10), 2435–2446.
United States v. City of Erie, Pennsylvania, 352 F. Supp. 2d 1105 (W.D. Pa. 2005).
UWUA Local 223 & The Detroit Edison Co., AAA Case No. 54–30–1746–87 (Apr. 17, 1991) (Lipson, Arb.)
Varden v. City of Alabaster, Alabama and John Cochran, U.S. District Court, Northern District of Alabama, Southern Division, 2:04-CV-0689-AR. 2004.
Vasich v. City of Chicago, 11 cv 4843 (N.D. Ill.) 2013.
White v. Village of Homewood, 628 N.E.2d 616 (Ill. App. 1993).
Wigdor, A. K., & Green, B. F. (1991). *Performance assessment for the workplace*. Washington, DC: National Academy Press.