FIELD MEASURES OF STRENGTH AND FITNESS PREDICT FIREFIGHTER PERFORMANCE ON PHYSICALLY DEMANDING TASKS

NORMAN D. HENDERSON
Department of Psychology Oberlin College

MICHAEL W. BERRY
Milwaukee Fire Department (Ret.), Milwaukee

TOMISLOV MATIC
Milwaukee Fire Department (Ret.), Milwaukee

Recruits from 9 consecutive fire academy classes were assessed on a battery of strength and endurance measures at Weeks 1, 7, and 14 of training. Regression analyses using Week 1 measures indicated that strength variables were the primary predictors of performance on physically demanding firefighting tasks assessed at the completion of training. Incremental validity was obtained with the addition of aerobic capacity, which produced more accurate performance distinctions among recruits with high strength levels. Results showed predictive validity and linearity throughout the upper range of strength and endurance levels, but evidence of an increased drop-off in performance for recruits with strength levels below the male 25th percentile. Structural equation modeling, a more powerful and comprehensive approach to validation than traditional regression, provided strong support for the construct validity of general strength and endurance as predictors of firefighter performance on physically demanding fire suppression and rescue tasks.

Physical ability selection tests for municipal firefighters rapidly shifted away from the use of measures of strength and endurance based on criterion or construct validation models, following an initially successful legal challenge to this approach (Berkman v. the City of New York, 1982). The federal trial court ruled that although the test was professionally developed, the research on which it was based measured general human ability rather than firefighting skills in particular. Assessments of muscular strength and general fitness were largely replaced by tests based on a content validity strategy using simulated fire suppression and rescue tasks. This change was probably due in part to the perception that such tasks would be more readily accepted as “job related” by courts than field tests derived from...
exercise physiology that were designed to assess strength and endurance. Shifting to a content validity strategy also relieves test developers from the obligation to carry out criterion-based validation studies, which, even when technically feasible, are expensive and pose formidable difficulties associated with testing incumbents on predictor measures. As a consequence, empirical validation studies of physical ability predictors for safety forces remain infrequent in the literature, despite concern about the paucity of this work raised by the editor of the *Journal of Applied Psychology* (Campbell, 1982).

Although criterion-based validation research on safety forces is relatively sparse, the importance of physical abilities in both job performance (e.g., Arvey, Nutting, & Landon, 1992; Blakley, Quinones, Crawford, & Jago, 1994; Hoffman, 1999; Hogan, 1991a; Sothmann, Gebhardt, Baker, Kastello, & Sheppard, 2004) and in the reduction of job-related injuries (e.g., Beaton, Murphy, Salazar, & Johnson, 2002; Harbin & Olson, 2005; Jackson, 1994; Stevenson, Weber, Smith, Dumas, & Albert, 2001) has received substantial recent attention.

Hogan (1991a) summarized 14 studies that correlated various physical ability measures with training time, ratings of job performance, and work simulations in a variety of physically demanding jobs. She found generally high validity coefficients, especially when the criterion measures involved work samples. Hogan also summarized the results of an unpublished meta-analysis of 10 mixed gender applicant samples (\(N = 2,064\)) by Lewis (1989), in which muscular strength was correlated with work simulations for physically demanding jobs. The corrected mean validity coefficient (\(\rho\)) was .82. In six samples (\(N = 1,740\)) correlating muscular endurance with work samples, Lewis obtained a \(\rho\) of .37. Shortly thereafter, Blakley et al. (1994) produced similar, although less striking, meta-analytic results (\(\rho = .55\)) for seven physically demanding jobs (\(N = 1,364\)) that correlated strength and work sample performance. As Hogan (1991a) amply demonstrated, however, within-gender validity coefficients tend to be considerably smaller than the high values found in mixed gender samples. Hogan provides a good overview of the gender–strength relationship as it pertains to test validation.

Hogan (1991b) has also demonstrated that physical performance test batteries demonstrate a fairly consistent structure across different occupational groups. A principal components analysis of each of the seven occupational groups yielded both a general strength component that accounted for a substantial proportion of test variance and an endurance component, accounting for less variance. The strength factor tended to encompass both static and dynamic assessments, as well as power and, to some degree, muscular endurance. Her endurance component is more difficult to define precisely because of the inconsistency in measurements
across studies. For example, only two of the seven studies directly assessed aerobic capacity.

*Problems With Simulation Tests for Entry-Level Firefighters*

Although now prevalent for entry-level selection, content-based assessments of physical abilities remain vulnerable to challenge. Issues related to the adequacy of underlying job analyses to tap relevant underlying physical abilities and the fidelity between responses required on the test and those required on the job are two of the more common general problems (e.g., Arvey, Nutting, et al. 1992; Hogan, 1991a). With respect to the specific job of firefighting, the use of simulation tests for entry-level selection poses several potentially serious problems.

First, fire academies train recruits on skills related to effective execution of many of the tasks used in preemployment screening, raising questions about the appropriateness of using some of those tasks for entry-level selection (Society for Industrial and Organizational Psychology [SIOP], 2003, p. 23).

Second, performance on some simulated fire suppression and rescue tasks can improve substantially with nominal training and practice or task familiarization. To reduce score variance due to these factors, test developers often provide training manuals to applicants, and some employers hold orientation sessions for several weeks before testing. Applicants frequently have access to test facilities long before scheduled test dates and sometimes receive supervised practice at these facilities. These pretest activities place a heavy demand on municipalities’ resources and simultaneously make the employer vulnerable to applicant claims of unequal access to pretesting benefits. The high costs of test preparation and testing with simulated firefighting and rescue tasks are often passed on to candidates in terms of substantial examination fees, creating economic barriers and reducing the size of applicant pools and presenting another mechanism for potential adverse impact.

Third, content-based validation strategies require that effort levels, as reflected in factors such as weight of objects, leverage, distance, repetitions, and possibly time limits, must be established for each simulated task in the selection battery, despite the lack of any broadly accepted procedural methods for doing so. Because fire energy levels tend to increase as a function of the cube of time elapsed (Thomas, 1995), greater speed of successful task completion is nearly always associated with increased fire suppression and rescue success. But, because setting high speed and effort levels for simulation tasks increases the probability of a challenge to the selection procedure, there are incentives to set task effort levels
low, sometimes below those routinely encountered in fire suppression and rescue.

Fourth, simulation tests usually stipulate numerous procedural infractions that can result in penalties or disqualification, producing substantial score variance not attributable to physical skills and abilities. Additional extraneous score variance is generated in timed, multitask simulations where part of the time-to-completion involves low-effort activities, including between-task travel time, during which test-takers risk penalties for moving too fast.

Fifth, practical limits on the fidelity of some fire simulation tasks can lead to a discrepancy in underlying skills and abilities used to complete simulated and actual job tasks. At a fire scene, rescue victims do not have handles, swinging room for axes and sledges is often constrained, and the substantial use of leg muscles while simulating ceiling breach and pull activities exacts a heavy oxygen cost when using self-contained breathing apparatus (SCBA), to cite three examples. Simulated tasks rarely replicate actual structural fire environments that stress cardiorespiratory capability of firefighters (Sothmann, Landy, & Saupe, 1992).

Sixth, multitask simulations scored on the basis of a single completion time score often have greater job fidelity than measuring each task in isolation, but they are vulnerable to a “weakest link” challenge to validity. A successful challenge of a single task element in the series, because of its lack of job fidelity or its placement in the sequence, may result in a successful challenge to the content validity of the entire simulation.

Finally, in the absence of criterion-based validation studies of firefighter simulation tests, it is difficult to assess the benefits of this selection method relative to its considerable administration cost. The problem is exacerbated by the tendency for many content-based tests to be scored on a pass/fail basis, rather than as continuous scores to be combined with other elements in the selection process. When the cut-off scores are set low enough to allow very high passing rates, the physical test may contribute little to the predictive validity the overall selection procedure. High pass rates preclude any meaningful subsequent criterion-based validation research, even when continuous performance scores are retained, because high physical ability applicants can usually perform at a level well below their true capacity and still qualify.

We demonstrate here that easily administered assessments of strength and endurance show empirically strong relationships with success of firefighter recruits in executing fire suppression tasks following training. These strength and endurance assessments can be highly standardized and avoid most problems associated with job-content selection procedures. They may also prove to have higher reliability and validity coefficients as initial screening devices for fire service than many content-based fire scene simulation tests. We then apply a construct validity approach to the
data, similar to that used in the development of physical ability tests for police officers (Arvey, Landon, Nutting, and Maxwell [1992]). The results provide strong evidence for construct validity of strength and endurance as predictors of demanding physical task performance in firefighting and rescue.

Method

Subjects

Data were obtained for all members of nine consecutive fire recruit classes hired by the city of Milwaukee between 1992 and 1998. Class sizes ranged from 23 to 46 and totaled 287 male and 19 female recruits. Each recruit class received a 16-week training program. All recruits were assessed on physical ability and anthropometric measures during weeks 1, 7, and 14 of fire academy training. Because nearly all male applicants who took firefighter physical ability entry-level tests in the city during this period received passing scores, male members of these nine academy classes are reasonably representative of their respective applicant pools, with respect to physical abilities. Female recruits represented approximately the top scoring quarter of female applicants who took the physical ability entry-level tests.

Anthropometric Measures

Height, weight, and skinfold measurements were obtained for all recruits on three occasions, as indicated above. A percent body fat estimate was determined based on abdominal or suprailliac skinfold measurements using a Slim Guide® or Accu-Measure® caliper in conjunction with age and gender based conversion tables. Lean body mass estimate was then computed from the percent body fat estimate and body weight.

Physical Ability Assessments

Upper-body strength was assessed using a bench press, a lat pull-down, and a grip strength measure, and a 1-minute sit-up test was used to measure abdominal strength and endurance. Aerobic endurance, the maximal capacity to take in and utilize oxygen, was assessed with a 5-minute step test.

Bench press. Strength testing equipment for the bench press was the bench press station of a Universal multiStation machine or a Badger Magnum multiStation machine. Recruits assumed a supine position on the bench with feet on the floor. The recruit began the testing procedure by performing one set of 15 repetitions with a low effort weight. The testers’
expertise was used to determine the initial warm-up set and the increases between sets based on the perceived strength abilities of the individual recruit (size and gender). An initial weight of approximately 32 kg was used for males and approximately 18 kg for females, followed by a 9–14 kg increase between the initial warm-up set and the first attempt. Using a standard grip that aligns forearms perpendicular to the floor, the recruit pushed smoothly until the weight was raised to full extension. The weight was lowered to the starting position and the action repeated. The initial set was followed immediately by consecutive and progressively heavier one repetition attempts, typically 4.5–14 kg, based on the recruits’ previous response, until failure was reached, usually between four and six attempts.

The last successfully lifted weight was used as the recruit’s test weight for the first test period. After a 5–6 minutes recovery period, the recruit completed as many repetitions as possible with the test weight. When the number of repetitions exceeded one, the total number of repetitions of the test weight was converted to a one repetition maximum estimate (1-RMEst) using Epley’s formula: \(1\text{-RMEst} = \text{test weight} + .033 (\text{test weight} \times \# \text{repetitions})\), cited in Ware, Clemens, Mayhew, and Johnston (1995).

**Lat pull-down.** Strength testing equipment for the lat pull-down was the lat pull-down station of a Badger Magnum multistation machine or a Badger Magnum single-station machine. Recruits assumed a seated position with seat or restraining pad adjusted to the proper position. The recruit reached up and grasped the lat bar, using a shoulder width grip with palms up, and pulled the bar down smoothly until it touched the sternum or the hands touched the sides of the chest. The bar was then raised, lowering the weight, until the arms were extended fully. The 1-RMEst testing procedure described for the bench press (i.e., 15 low-effort reps, increasing weight until failure, recovery, reps with test weight until failure) and the 1-RMEst conversion formula were also used to obtain the lat pull-down 1-RMEst score at the first test period.

For both bench press and lat pull-down, a similar procedure was followed at the 7th and 14th week strength retest; a light 15 rep warm-up set using 50% of the previous 1-RMEst was done first, followed by a recovery period. Following the recovery period, a 10-RM test weight (.759 of the previous 1-RMEst, based on Epley’s formula) was lifted to failure by the recruit to establish a current 1-RMEst. The 10-RM weight represents the weight recruits should be able to lift 10 times based on their previous strength test 7 weeks earlier. Completing more or less reps than 10 indicated either a strength increase or a strength loss.

**Grip strength.** Strength testing equipment for grip strength was a Baseline® or Rolyan® hydraulic hand dynamometer adjusted to the
approximate size of an axe handle. Each recruit received two trials with their dominant hand, with the better of the two attempts recorded.

**Step test.** The recruits’ maximal aerobic power ($V_{O_2max}$) was assessed using a 5-minute, 90-steps-per-minute pace step test, sometimes referred to as the U.S. Forestry Step test, (Davis & Wilmore, 1990). Conversion tables for estimating relative $V_{O_2max}$ in ml/[kg minute] from gender, body weight, age, and pulse rate can be found in Jacobs (1990). We use the term $V_{O_2max}$ in the text to refer to relative $V_{O_2max}$ expressed in terms of milliliters of oxygen per kilogram of body weight per minute. Expressing $V_{O_2max}$ as a function of body weight allows a direct comparison of individuals independent of body size. $V_{O_2max}$ is variously referred to as aerobic power, aerobic fitness, or maximal oxygen intake per kg of body weight. Absolute $V_{O_2max}$ (Abs $V_{O_2max}$), expressed in liters per minute (L/minute) is sometimes referred to as aerobic capacity—the total capacity of the cardiorespiratory systems, or maximal oxygen consumption in L/minute, was calculated from relative $V_{O_2max}$ and body weight.

**Sit-ups test.** The recruits’ general muscular endurance was assessed using the YMCA’s 1-Minute Timed Sit-Ups test (Golding, Myers, & Sinning, 1982). The recruits were on their back for the sit-ups, with knees bent at about a 90-degree angle and fingers laced behind head. On command, the recruit touched right elbow to left knee and returned to starting position, then touched left elbow to right knee, continuing this alternating cycle for one minute. The sit-ups score was the total number of correctly performed sit-ups in 60 seconds.

**Criterion Performance Measures**

Performance assessments were made midway through fire academy training and at the end of training in week 15 or 16. The roof ladder placement exercise was practiced at least once each week and the chopping exercise practiced daily throughout academy training. Training and assessment programs varied slightly from class to class, hence not all classes were evaluated on all the measures described below. Using physical ability measures in conjunction with end-of-training criterion tests with academy recruits still on probationary status substantially reduces motivational confounds often encountered when using postprobationary incumbents.

**One-Person Roof Ladder Placement (All Academy Classes)**

Recruits shoulder a 6.1 m, 26.3 kg roof ladder approximately 8 m from a (9.2 m) wall ladder anchored to a mock two story residential structure. The recruit carries the roof ladder to the structure, opens the hooks of the
roof ladder and raises it to a vertical position alongside the wall ladder. The recruit ascends the wall ladder to a point where he/she can grasp and place the roof ladder’s fourth rung (from the top) on their shoulder. The recruit ascends the wall ladder carrying the roof ladder to a position where he/she can lock-in to the ladder (using the leg opposite the shoulder carrying the roof ladder) with their shoulders at or above the eaves of the roof. The roof ladder is then placed onto the roof on its side-rail (beam) and pushed up until the hooks of the ladder clear the top edge of the roof (ridge). The roof ladder is then laid flat to engage the hooks. Timing stops when the recruit completes the ladder placement, unlocks from the wall ladder, and places a hand on the roof ladder in position to transfer from the wall ladder to the roof ladder. Recruit is in full turn-out gear, including spanner belt with axe, but without SCBA, for a total weight of 18.4 kg.

**Chopping Exercise Ratings (Five Academy Classes)**

As a warm-up, recruits first chopped with the 3.6 kg fire axe for 1.5 minutes while standing at ground level. Next, following a short rest period, the recruit ascended a roof ladder (a distance of 3 m) that was secured to a 45-degree incline board and chopped for 3 minutes (mid-training session) or 4 minutes (end of training session) from a kneeling position on the right side. After another short rest period, the recruit chopped for 3 minutes from a kneeling position on the left side. Four or five academy instructors evaluated the recruits on power and endurance for both the left and right chopping positions. Power was evaluated based on acceleration of the ax head, observed depth of cut, sound (a powerful stroke has a full sharp sound on the incline board, as opposed to a dull thud when the ax head is “dropped” instead of accelerating), and a visible bounce of the log in its holder when a powerful blow is administered. Endurance was based on the ability of the recruit to maintain an acceptable pace at their demonstrated power from beginning to end. Average instructor rating was used to establish power and endurance scores for left and right chopping positions. On infrequent occasions when there was an outlier rating among instructors, the ratings were discussed and sometimes modified before averaging. Scores for left and right positions were later averaged for both power and endurance ratings. Individual rater scores were not retained, so interrater reliability coefficients could not be computed. Mid- versus end-of-training test–retest reliabilities were available.

**Combat Test (Two Academy Classes)**

Except for minor differences in distances between events, the combat test was nearly identical to that used in the U.S. Combat Challenge
competition, described in detail at the www.firefighter-challenge.com Web site. Recruits in full turnout gear, including SCBA (29.3 kg total weight), completed a course that consisted of five events: a five floor high rise stair climb with a hose pack weighing 20.4 kg; a hose hoist evolution in which recruit uses a hand-over-hand motion to hoist a 20.4 kg hose pack to the fifth floor of the training tower, pulling smoothly to ensure that the load does not become tangled; a forcible entry evolution using a 4 kg sledge in a simulation that reflects the biomechanical activity of chopping; a hose advance exercise in which the recruit picks up the nozzle of a fully charged 4.5 cm hose line, moves it 42.7 m, opens the nozzle for one second and places the nozzle in a box; a simulated victim rescue, dragging a 75.2 kg dummy backwards 30.5 m. Among the five events participants travel a total of approximately 52 m and descend five floors. Instructors scoring the events were certified by the test developers (ARA Human Factors). Total time to complete all five events is used as the criterion measure.

Training Rank Assessment (Five Academy Classes)

This assessment procedure was developed and implemented by academy instructors several years before this study was conceived. At the end of the training class, the team of four or five fire academy instructors evaluated the overall performance of each recruit on the fire suppression and rescue tasks taught at the academy. Instructors then chose, by consensus, the five overall best and five poorest performing recruits in that academy class. Individual instructor ratings were not recorded for this measure, thus we were limited to using a three-step rating scale for each recruit, with approximately 15% low, 70% middle, and 15% high scores assigned.

Results

Physical Ability Assessments

We first extend Hogan’s (1991b) Table 6 by adding an orthogonal Principal Components analysis of our physical performance variables, assessed both at the beginning and at the end of academy training. The results of our two analyses, along with total group means, SDs, and standardized improvement scores are shown in Table 1. Our measures produced a strength component and an endurance/fitness component on both test occasions that are quite similar to those reported by Hogan. For convenience, we use the terms “strength” and “endurance” throughout the remainder of the paper. Our strength component is more precisely labeled “upper body isotonic strength” based on the underlying measures. Similarly,
<table>
<thead>
<tr>
<th>Measure</th>
<th>Week 1 assessments</th>
<th>Week 14 assessments</th>
<th>Improvement&lt;sup&gt;b&lt;/sup&gt;</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (SD)</td>
<td>Factor 1</td>
<td>Factor 2</td>
<td></td>
</tr>
<tr>
<td>Bench press (kg)</td>
<td>98.4 (27.2)</td>
<td>.83</td>
<td>.25</td>
<td>.47</td>
</tr>
<tr>
<td>Lat pull-down (kg)</td>
<td>105.2 (17.7)</td>
<td>.85</td>
<td>.22</td>
<td>.88</td>
</tr>
<tr>
<td>Grip strength (kg)</td>
<td>64.4 (10.0)</td>
<td>.81</td>
<td>−.15</td>
<td>.14</td>
</tr>
<tr>
<td>Sit-ups (60-sec total)</td>
<td>46.2 (7.6)</td>
<td>.16</td>
<td>.75</td>
<td>.74</td>
</tr>
<tr>
<td>Relative Vo&lt;sub&gt;2max&lt;/sub&gt; (ml/[kg minute])</td>
<td>42.6 (5.7)</td>
<td>−.01</td>
<td>.76</td>
<td>1.58</td>
</tr>
<tr>
<td>Percent body fat</td>
<td>13.7 (4.7)</td>
<td>−.11</td>
<td>−.80</td>
<td>−.42</td>
</tr>
<tr>
<td>Percentage of variance</td>
<td>35.0</td>
<td>31.9</td>
<td>37.5</td>
<td>25.0</td>
</tr>
</tbody>
</table>

<br><br><sup>a</sup>Principle Components Analysis with Varimax rotation. Total N = 306 (287 men, 19 women).

<sup>b</sup>(M<sub>wk1</sub> − M<sub>wk14</sub>)<i>/SD</i> <sub>wk1</sub>. 

### TABLE 1

Rotated Factor Loadings of Strength and Endurance Predictor Variables<sup>a</sup>
our endurance component involves aerobic fitness relative to body weight, muscular endurance, and general systemic fitness involving large muscles and is sometimes called stamina or sustainable fitness. Over the course of academy training, recruits improved significantly on all strength and endurance measures (all \( p < .0001 \) with paired sample \( t \)-tests).

Because there are large gender differences on many physical measures, total sample correlation coefficients are influenced by both the proportion of women in the total sample and by the size of the gender difference on each variable. Thus, although mixed gender correlations are appropriate to assess validity coefficients for a given applicant pool, they are less useful for providing standard benchmarks for comparisons across studies than same-sex or pooled within-sex correlations. Because our sample is 94% male we have chosen to present most of our correlation analyses for men only. In most cases the mixed-gender correlations are slightly larger than those reported for men alone. For some criterion performance measures, the number of women for whom data are available are too small for any meaningful analysis.

Table 2 shows the correlations among the various strength and endurance measures assessed at the beginning and at the end of training. Test–retest reliability coefficients across the 14-week interval are shown on the diagonal in Table 2. Mixed-gender coefficients were notably larger for the lat pull (.80 vs. .75) and grip strength (.77 vs. .69) measures and only slightly larger for the remaining measures. Aerobic capacity measures showed lower test–retest reliability than strength measures. The other notable result in Table 2 is the shift in \( V_{O2\max} \)--strength correlations from low positive at the beginning of training to low negative at the end of training. Among men, rate of improvement was uncorrelated for strength and aerobic endurance measures (\( r = .02 \)).

**Performance Criterion Measures**

The criterion measures were originally collected for training purposes within each class and not with a view towards pooling data across years. As a consequence, there were some procedural differences in criterion assessment from class to class as well as some turnover in academy instructors. These changes led to some between-class mean criterion differences that were the result of procedural or measurement variations unrelated to actual recruit differences in performance. Some data conditioning was thus required before pooling of the nine recruit classes for the validation study. We chose not to standardize criterion measures within class using \( z \)-scores because of the loss of all true between- and within-class variance. When there are multiple groups and distinct nonrelevant sources of between-group variation, a preferred method is to generate correlation
TABLE 2

**Strength and Endurance Correlations for Male Recruits at Week 1 and Week 14**

<table>
<thead>
<tr>
<th>Strength/endurance measure</th>
<th>Rel Vo$_2$</th>
<th>Abs Vo$_2$</th>
<th>bench press</th>
<th>Lat pull</th>
<th>Grip strength</th>
<th>Composite</th>
<th>Sit-ups</th>
<th>LBM</th>
<th>% Fat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative Vo$_{2\text{max}}$ (ml/[kg minute])</td>
<td>.48</td>
<td>.58</td>
<td>.18</td>
<td>.14</td>
<td>-.08</td>
<td>.10</td>
<td>.31</td>
<td>-.16</td>
<td>-.46</td>
</tr>
<tr>
<td>Absolute Vo$_{2\text{max}}$ (L/minute)</td>
<td>.50</td>
<td><strong>.64</strong></td>
<td>.30</td>
<td>.35</td>
<td>.29</td>
<td>.39</td>
<td>-.02</td>
<td>.65</td>
<td>.16</td>
</tr>
<tr>
<td>Bench press (kg)</td>
<td>-.20</td>
<td>.10</td>
<td><strong>.90</strong></td>
<td>.55</td>
<td>.34</td>
<td>.77</td>
<td>.18</td>
<td>.35</td>
<td>-.08</td>
</tr>
<tr>
<td>Lat pull-down (kg)</td>
<td>-.21</td>
<td>.20</td>
<td>.64</td>
<td><strong>.75</strong></td>
<td>.35</td>
<td>.74</td>
<td>.18</td>
<td>.51</td>
<td>-.08</td>
</tr>
<tr>
<td>Grip strength (kg)</td>
<td>-.16</td>
<td>.21</td>
<td>.34</td>
<td>.43</td>
<td><strong>.69</strong></td>
<td>.66</td>
<td>-.02</td>
<td>.46</td>
<td>.10</td>
</tr>
<tr>
<td>Composite strength</td>
<td>-.23</td>
<td>.16</td>
<td>.84</td>
<td>.91</td>
<td>.40</td>
<td><strong>.90</strong></td>
<td>.22</td>
<td>.49</td>
<td>-.07</td>
</tr>
<tr>
<td>Sit-ups (total)</td>
<td>.04</td>
<td>-.23</td>
<td>.17</td>
<td>.17</td>
<td>.00</td>
<td>.19</td>
<td><strong>.77</strong></td>
<td>-.19</td>
<td>-.33</td>
</tr>
<tr>
<td>Lean body mass (kg)</td>
<td>-.36</td>
<td>.56</td>
<td>.36</td>
<td>.51</td>
<td>.44</td>
<td>.48</td>
<td>-.18</td>
<td><strong>.97</strong></td>
<td>.36</td>
</tr>
<tr>
<td>Percent body fat</td>
<td>-.29</td>
<td>.30</td>
<td>-.04</td>
<td>-.02</td>
<td>-.03</td>
<td>-.04</td>
<td>-.35</td>
<td>.29</td>
<td><strong>.85</strong></td>
</tr>
</tbody>
</table>

*Week 1 correlations above diagonal and week 14 correlations below the diagonal. Diagonal values are test–retest correlations. $N = 287$. 
coefficients from a pooled within-group variance–covariance matrix. Although this procedure solves the problem of class-to-class variation in criterion assessment methods, it also restricts true between-class variance in some physical ability measures of recruits, which were selected under some widely differing physical selection procedures and selection ratios. For example, mean class-to-class differences in the three strength measures of male recruits were all substantial (10–22% of total variance, all \( p < .001 \)). Excluding this variance among classes has the effect of restricting range and thus reducing the correlations among some measures. Although range restriction corrections can be applied to the within-class correlations, the resulting coefficients cannot be used with the powerful full information maximum likelihood estimates (FIML) used in this paper.

Instead, using the data from all male recruits, we carried out a separate analysis of covariance for each criterion variable, using the strength and endurance measures as covariates. No significant between-class differences were found for training rank or for the combat test time, so raw scores were used in all subsequent analyses with these two criterion variables. Covariate-adjusted means for the two axe performance ratings and for roof ladder time differed significantly across academy classes. The nine covariate-adjusted class means were then equated by adding the appropriate constant (i.e., grand mean–class mean) to all recruit scores within an individual class. For these three criterion variables, recruit raw scores were thus replaced by their class-equated scores for use in subsequent analyses. As shown below, the correlation coefficients obtained with these class-equated scores are nearly identical to those calculated from the pooled raw score within-class variance–covariance matrix, which is then corrected for range restriction. Correlations between class-equated scores and their raw scores are \( r = .93 \) (roof ladder), \( r = .90 \) (axe power), and \( r = .85 \) (axe endurance). A composite measure was also created from the sum of the \( z \)-transformed bench press, lat pull-down, and grip strength measures. This strength composite score was rescaled as a \( T \)-score, with a mean of 50 and \( SD \) of 10. An endurance composite \( T \)-score was also created, using the sum of \( z \)-scores for \( Vo_2_{\text{max}} \) and sit-ups.

Because each team of instructors chose by consensus the best and poorest performing five recruits at the end of the training class, we were limited to using a three-step rating scale for each recruit. We converted these categorizations into a +1, 0, −1 scale, representing the top five recruits, the middle group, and bottom five recruits in each class. Because selectivity for inclusion in top and bottom five varied as a function of class size, we substituted probit scores, corresponding to percentile cut-points required to choose top and bottom recruits in each class, in place of the +1 and −1 of the three-step scale. Results using the probit scores differed
trivially from those using the $+1$, $0$, $-1$ values, hence results using the simpler scale are reported here.

Because there is information loss resulting from the underlying continuous distribution of performance being broken into high, middle, and low categories, containing approximately 15%, 70%, and 15% of cases, respectively, there is some attenuation of correlations between training rank scores and other variables. We ran simulations to estimate the degree of attenuation of correlation coefficients caused by trichotomizing the training criterion variable at these cut-points. Results indicated that correlations based on continuous normal scores are approximately 11% larger than correlations obtained when the data are categorized into a 15%–70%–15% split. We did not correct for this estimated shrinkage in our observed correlations, but the phenomenon is noted in the relevant tables.

Our initial analyses employed traditional least squares multiple regression methods. We then moved to a more powerful and comprehensive analysis using FIML structural equation modeling.

Table 3 lists Pearson correlations between end-of-training performance, and strength, endurance, and other physical characteristics of male recruits assessed during week 1 at the academy. Excluding the high (.83) correlation between the axe power and axe endurance ratings, correlations among criterion variables average .54, with a range of .45–.59. Correlations are higher than those tabled when female recruits are included in the analysis, particularly among pairs of variables where both show large gender differences. As indicated above, for this and most other analyses we focus on only male recruits because full sample correlations are influenced by the varying proportion of women included in each pair of variables.

Shown in parentheses in Table 3 are the raw score correlations based on pooled within-class variance–covariance matrices for the three criterion measures that showed significant between-class mean differences. These coefficients are very close to our covariate-adjusted coefficients and are nearly identical (mean absolute difference = .013) to the covariate-adjusted coefficients after they are corrected for range restriction.

Strength measures are consistently high predictors of criterion performance for all but the axe endurance rating. Relative $V_o^{2max}$ and number of sit-ups generally show lower validity coefficients than strength measures. Among anthropometric measures, height is a significant predictor of all criterion measures and lean body mass (LBM) is significant for all but axe endurance. The fire academy strength and endurance program was successful. Mean week 1 scores shown in Table 3 increased to 51.7 ($V_o^{2max}$) 11.3% (percent body fat), 113.9 (bench press kg), and 122.5 (lat pull-down kg), all significant at $p < .001$.

Table 4 shows the standardized regression weights for the strength composite measure, $V_o^{2max}$ and number of sit-ups obtained from multiple
TABLE 3

*Correlations Among Predictor, Criterion, and Subject Variables of Male Recruits*

<table>
<thead>
<tr>
<th>Measure</th>
<th>Mean (SD)</th>
<th>Training*</th>
<th>Ax power*</th>
<th>Ax endurance*</th>
<th>Combat test*</th>
<th>Roof ladder*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative VO₂max (ml/[kg minute])</td>
<td>42.7 (5.7)</td>
<td>.05</td>
<td>.33 (.35)</td>
<td>.40 (.41)</td>
<td>.15</td>
<td>-.19 (-.18)</td>
</tr>
<tr>
<td>Absolute VO₂max (L/minute)</td>
<td>3.7 (0.6)</td>
<td>.24</td>
<td>.32 (35)</td>
<td>.20 (.21)</td>
<td>.43</td>
<td>-.22 (-.22)</td>
</tr>
<tr>
<td>Bench press (kg)</td>
<td>101.5 (25.9)</td>
<td>.33</td>
<td>.21 (.21)</td>
<td>.08 (.08)</td>
<td>.33</td>
<td>-.43 (-.39)</td>
</tr>
<tr>
<td>Lateral pull-down (kg)</td>
<td>107.4 (16.2)</td>
<td>.34</td>
<td>.40 (.43)</td>
<td>.25 (.28)</td>
<td>.62</td>
<td>-.47 (-.46)</td>
</tr>
<tr>
<td>Grip strength (kg)</td>
<td>65.4 (9.1)</td>
<td>.30</td>
<td>.13 (.12)</td>
<td>.04 (.01)</td>
<td>.50</td>
<td>-.31 (-.26)</td>
</tr>
<tr>
<td>Composite strength</td>
<td>51.3 (8.8)</td>
<td>.41</td>
<td>.31 (.32)</td>
<td>.14 (.15)</td>
<td>.59</td>
<td>-.50 (-.47)</td>
</tr>
<tr>
<td>Sit-ups (total)</td>
<td>46.3 (7.5)</td>
<td>.20</td>
<td>.28 (.29)</td>
<td>.28 (.30)</td>
<td>.31</td>
<td>-.31 (-.31)</td>
</tr>
<tr>
<td>Endurance composite</td>
<td>50.3 (9.8)</td>
<td>.16</td>
<td>.38 (.37)</td>
<td>.42 (.43)</td>
<td>.28</td>
<td>-.31 (-.31)</td>
</tr>
<tr>
<td>Height (cm)</td>
<td>180.5 (6.4)</td>
<td>.29</td>
<td>.34 (.35)</td>
<td>.21 (.20)</td>
<td>.48</td>
<td>-.20 (-.20)</td>
</tr>
<tr>
<td>Lean body mass (kg)</td>
<td>74.4 (8.1)</td>
<td>.32</td>
<td>.19 (.21)</td>
<td>-.03 (-.03)</td>
<td>.52</td>
<td>-.19 (-.19)</td>
</tr>
<tr>
<td>Percent body fat</td>
<td>13.3 (4.5)</td>
<td>-.10</td>
<td>-.31 (-.31)</td>
<td>-.38 (-.38)</td>
<td>-.17</td>
<td>.25 (25)</td>
</tr>
<tr>
<td>Age (years)</td>
<td>26.1 (4.7)</td>
<td>-.21</td>
<td>-.39 (-.41)</td>
<td>-.35 (-.37)</td>
<td>-.03</td>
<td>.22 (20)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Criterion correlations</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Training (consensus rank ×10)</td>
<td>20.4 (5.1)</td>
<td>.54</td>
<td>.45</td>
<td>.61</td>
<td>-.58</td>
<td></td>
</tr>
<tr>
<td>Axe power rating (×10)</td>
<td>37.4 (5.5)</td>
<td>.72b</td>
<td>.83</td>
<td>.63</td>
<td>-.59</td>
<td></td>
</tr>
<tr>
<td>Axe endurance rating (×10)</td>
<td>37.6 (4.8)</td>
<td>.67</td>
<td>.56</td>
<td>-.51</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combat test speed (sec⁻¹)i</td>
<td>229.4 (15.1)</td>
<td>.76</td>
<td>.76</td>
<td>-.49</td>
<td></td>
<td>.75</td>
</tr>
<tr>
<td>Roof ladder time</td>
<td>48.7 (7.8)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*N = 167. Correlations ≥ .19 significant at the p < .01 level, (2-tail). Observed correlations of training ratings are attenuated due to categorization of the criterion variable. Approximate corrected correlation = 1.11 × observed r. A composite axe power and endurance measure correlates slightly above the average of the individual axe power and axe endurance correlations with predictor measures.

*bSpeed to completion (sec⁻¹). N = 60. Correlations ≥ .26 are significant at the p < .05 level; Correlations ≥ .33 are significant at the p < .01 level (2-tail).

*cN = 287. All correlations are significant at the p < .001 level, (2-tail).

*dT-scores, based on all 306 recruits from the nine classes.

*eCorrelations within parentheses are based on pooled within-class raw scores.

*fPooled z-scores of Relative VO₂ max and sit-up scores.

*gCorrelation increases by approximately .03 when corrected for attenuation related to loss of between-class mean strength differences.

|hTest-retest reliability coefficients (week 7 vs. week 14 scores) are shown on the diagonal.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Mean (SD)</th>
<th>Training*</th>
<th>Ax power*</th>
<th>Ax endurance*</th>
<th>Combat test*</th>
<th>Roof ladder*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative VO₂max (ml/[kg minute])</td>
<td>42.7 (5.7)</td>
<td>.05</td>
<td>.33 (.35)</td>
<td>.40 (.41)</td>
<td>.15</td>
<td>-.19 (-.18)</td>
</tr>
<tr>
<td>Absolute VO₂max (L/minute)</td>
<td>3.7 (0.6)</td>
<td>.24</td>
<td>.32 (35)</td>
<td>.20 (.21)</td>
<td>.43</td>
<td>-.22 (-.22)</td>
</tr>
<tr>
<td>Bench press (kg)</td>
<td>101.5 (25.9)</td>
<td>.33</td>
<td>.21 (.21)</td>
<td>.08 (.08)</td>
<td>.33</td>
<td>-.43 (-.39)</td>
</tr>
<tr>
<td>Lateral pull-down (kg)</td>
<td>107.4 (16.2)</td>
<td>.34</td>
<td>.40 (.43)</td>
<td>.25 (.28)</td>
<td>.62</td>
<td>-.47 (-.46)</td>
</tr>
<tr>
<td>Grip strength (kg)</td>
<td>65.4 (9.1)</td>
<td>.30</td>
<td>.13 (.12)</td>
<td>.04 (.01)</td>
<td>.50</td>
<td>-.31 (-.26)</td>
</tr>
<tr>
<td>Composite strength</td>
<td>51.3 (8.8)</td>
<td>.41</td>
<td>.31 (.32)</td>
<td>.14 (.15)</td>
<td>.59</td>
<td>-.50 (-.47)</td>
</tr>
<tr>
<td>Sit-ups (total)</td>
<td>46.3 (7.5)</td>
<td>.20</td>
<td>.28 (.29)</td>
<td>.28 (.30)</td>
<td>.31</td>
<td>-.31 (-.31)</td>
</tr>
<tr>
<td>Endurance composite</td>
<td>50.3 (9.8)</td>
<td>.16</td>
<td>.38 (.37)</td>
<td>.42 (.43)</td>
<td>.28</td>
<td>-.31 (-.31)</td>
</tr>
<tr>
<td>Height (cm)</td>
<td>180.5 (6.4)</td>
<td>.29</td>
<td>.34 (.35)</td>
<td>.21 (.20)</td>
<td>.48</td>
<td>-.20 (-.20)</td>
</tr>
<tr>
<td>Lean body mass (kg)</td>
<td>74.4 (8.1)</td>
<td>.32</td>
<td>.19 (.21)</td>
<td>-.03 (-.03)</td>
<td>.52</td>
<td>-.19 (-.19)</td>
</tr>
<tr>
<td>Percent body fat</td>
<td>13.3 (4.5)</td>
<td>-.10</td>
<td>-.31 (-.31)</td>
<td>-.38 (-.38)</td>
<td>-.17</td>
<td>.25 (25)</td>
</tr>
<tr>
<td>Age (years)</td>
<td>26.1 (4.7)</td>
<td>-.21</td>
<td>-.39 (-.41)</td>
<td>-.35 (-.37)</td>
<td>-.03</td>
<td>.22 (20)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Criterion correlations</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Training (consensus rank ×10)</td>
<td>20.4 (5.1)</td>
<td>.54</td>
<td>.45</td>
<td>.61</td>
<td>-.58</td>
<td></td>
</tr>
<tr>
<td>Axe power rating (×10)</td>
<td>37.4 (5.5)</td>
<td>.72b</td>
<td>.83</td>
<td>.63</td>
<td>-.59</td>
<td></td>
</tr>
<tr>
<td>Axe endurance rating (×10)</td>
<td>37.6 (4.8)</td>
<td>.67</td>
<td>.56</td>
<td>-.51</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combat test speed (sec⁻¹)i</td>
<td>229.4 (15.1)</td>
<td>.76</td>
<td>.76</td>
<td>-.49</td>
<td></td>
<td>.75</td>
</tr>
<tr>
<td>Roof ladder time</td>
<td>48.7 (7.8)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### TABLE 4

**Standardized Regression (Beta) Weights of Strength and Endurance Measures Obtained from Male Recruits at Week 1 and at Week 14 of Fire Academy Training**

<table>
<thead>
<tr>
<th>Performance criterion (Week 15)</th>
<th>Beta weight</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Strength composite</td>
</tr>
<tr>
<td>Roof ladder time ($N = 287$)</td>
<td></td>
</tr>
<tr>
<td>Predictors measured week 1</td>
<td>$-.45^c$</td>
</tr>
<tr>
<td>Predictors measured week 14</td>
<td>$-.53^c$</td>
</tr>
<tr>
<td>Axe rating ($N = 167$)</td>
<td></td>
</tr>
<tr>
<td>Predictors measured week 1</td>
<td>$+.15^c$</td>
</tr>
<tr>
<td>Predictors measured week 14</td>
<td>$+.27^c$</td>
</tr>
<tr>
<td>Combat test speed ($N = 60$)</td>
<td></td>
</tr>
<tr>
<td>Predictors measured week 1</td>
<td>$+.55^c$</td>
</tr>
<tr>
<td>Predictors measured week 14</td>
<td>$+.49^c$</td>
</tr>
<tr>
<td>Training consensus rank ($N = 167$)$^b$</td>
<td></td>
</tr>
<tr>
<td>Predictors measured week 1</td>
<td>$+.39^c$</td>
</tr>
<tr>
<td>Predictors measured week 14</td>
<td>$+.42^c$</td>
</tr>
</tbody>
</table>

$^a$All multiple correlations significant at $p < .001$.

$^b$Multiple $R$ approximately .48 when corrected for categorization of criterion variable.

$^c$Beta weight significant ($p < .01$), all others, $p > .05$.

Regression analysis. Because height, weight, and body fat would not normally be used as predictors in an employment setting, these measures are not included in this regression. For these and subsequent analyses, axe power and axe endurance ratings were pooled into a single axe rating measure because the correlation between the two ratings was .83. For consistency across criterion measures, all three predictors were entered into the regression, even though in some cases one or more of the predictors is not statistically significant at the .05 level. Because all predictor variables were measured at the beginning of academy training and again during week 14 of training, two regression equations were computed for each criterion measure. Additional analyses, not shown, were run in which the three individual strength measures were entered in lieu of the equally weighted composite. The multiple correlations and the regression weights of $V_{O2max}$ and sit-up variables departed slightly from those obtained with the *composite strength* measure. For all criterion variables, we looked for interactions among predictors using cross-product variables. None were statistically significant, with $p$ levels from .15 to .87. We also checked for nonlinearity using quadratic forms of the predictors. We found a significant curvilinearity effect for the strength variable in predicting both training rank ($p = .036$) and roof ladder time ($p < .001$). The issue of
prediction-criterion slope changes across strength levels is examined in more detail below.

Evident from Table 4 is the dominant role of strength in nearly all regression equations. Although recruits showed significant and sometimes substantial increases in strength and endurance measures across the 14 weeks of training (shown in Table 1) regression equations for beginning- and end-of-training measures were remarkably similar. The remaining analyses thus focus on beginning-of-training measures because we are primarily interested in the use of strength and endurance measures as predictors of subsequent firefighter physical task performance.

We looked for gender bias effects in the predictor variables by examining the significance of gender and interaction terms involving gender, following entry of the three predictors into each regression equation. This procedure tests Cleary’s (1968) model of test fairness. The process was then repeated, entering height and LBM into the regression with the strength and endurance predictor variables before entering gender and gender interaction variables.

There was no evidence of differential prediction by gender for axe rating, but women received lower training rank scores than predicted by the strength/endurance (S/E) measures \( p = .003 \), and the effect persisted when height and LBM were added to the equation. Women took significantly longer \( p < .0001 \) to complete the roof ladder evolution than was predicted by the strength and endurance predictors, with little change following addition of height and LBM to the equation. This result is not surprising because the standardized gender difference \( d = [M_{female} - M_{male}] / SD_{pooled} = 2.75 \) in roof ladder time was the largest gender effect observed among all predictor and criterion variables in the study. The combat test was excluded from gender analysis because data was available for only one female recruit.

Using the male sample, a regression analysis was computed to detect age bias effects. There was no evidence of age bias for training rank \( p = .112 \), combat test \( p = .428 \), or the roof ladder evolution \( p = .065 \), but age is a significant incremental predictor of the axe rating \( p < .0001 \) when added to strength and endurance measures. Older males receive lower ratings than predicted from their S/E scores, and the difference persists when height, percent body fat and LBM are added to the regression equation \( p = .0004 \).

Figure 1 is a scatterplot depicting the relationship between the strength composite variable assessed at the beginning of training and time to complete the roof ladder evolution at the end of training. The plotted linear regression line in Figure 1 is based on the 287 male recruits in the group. It is evident from the plotted points that the linear regression substantially underpredicts ladder completion times for subjects with the lowest
strength levels, a subgroup that includes all female recruits. Low strength recruits perform more poorly on the criterion measure than predicted from their strength levels.

Table 5 summarizes standardized (beta) weights, used to predict roof ladder time for each of the individual strength measures, when the strength, \( V_o2_{max} \), and sit-up measures are in the regression. With the possible exception of grip strength in the male-only sample, each of the strength measures is the substantive predictor of roof ladder performance, with small but significant incremental validity obtained by including \( V_o2_{max} \) and sit-ups measures. For example, for male recruits, multiple \( R \) increases from .50 with the composite strength variable alone, to .54 when \( V_o2_{max} \) and sit-up variables are added to the regression (\( R^2 \) increase \( p < .0001 \)). As indicated in Table 5, \( V_o2_{max} \) is usually marginally significant in incrementally predicting male performance when the sit-ups variable is removed from the regression. Although the composite strength measure produces
TABLE 5
Predictions of End-of-Training Roof Ladder Evolution Completion Times From
Week 1 Measures of Upper Body Strength, Relative Vo$_{2\text{max}}$, and Sit-Ups

<table>
<thead>
<tr>
<th>Strength measure</th>
<th>Vo$_{2\text{max}}$</th>
<th>Sit-ups</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>beta$^a$</td>
<td>p</td>
</tr>
<tr>
<td>Males Only (N = 287)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bench press</td>
<td>-.37</td>
<td>.07</td>
</tr>
<tr>
<td>lat pull-down</td>
<td>-.41</td>
<td>.08</td>
</tr>
<tr>
<td>Grip strength</td>
<td>-.31</td>
<td>.28</td>
</tr>
<tr>
<td>Composite</td>
<td>-.45</td>
<td>.09</td>
</tr>
<tr>
<td>Full sample (N = 306)$^e$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bench press</td>
<td>-.49</td>
<td>.05</td>
</tr>
<tr>
<td>lat pull-down</td>
<td>-.54</td>
<td>.07</td>
</tr>
<tr>
<td>Grip strength</td>
<td>-.47</td>
<td>.14</td>
</tr>
<tr>
<td>Composite$^f$</td>
<td>-.59</td>
<td>.08</td>
</tr>
</tbody>
</table>

$^a$Strength measure listed in column one.

$^b$Multiple $R$ based on the three predictors.

$^c$p < .0001 in all cases. All significance levels are two-tailed.

$^d$Vo$_{2\text{max}}$ significant ($p < .03$, with beta weights of -.12 to -.14) when sit-ups excluded from the regression.

$^e$Includes 19 female recruits. Inclusion of a quadratic strength function in the regression increases multiple $R$ by .02 to .04. Female mean and (SD): bench press 58.0 (11.1); lat pull-down 74.2 (9.3); strength composite 30.2 (3.5); Vo$_{2\text{max}}$ 40.8 (5.2); sit-ups 43.6 (9.7).

$^f$Pooled within-gender correlations among predictors: strength composite w. Vo$_{2\text{max}}$ $r = .10$; strength composite w. sit-ups $r = .22$; Vo$_{2\text{max}}$ w. sit-ups $r = .33$.

the highest validity coefficient, validity loss is slight when either bench press or lat pull-down is used as the sole strength assessment.

With the full sample of male and female recruits, there is a significant quadratic (curvilinear) effect ($p < .001$) for each of the strength measures predicting time to complete the roof ladder evolution. For the strength composite measure, an inverse function, ladder time = 1295 (strength composite $T$-score$^{-1}$) + 23, provided the best fit ($R = .67$, $p < .0001$) to the data. Further exploration of the relationship between strength and roof ladder time indicated that performance drop-off increases substantially for recruits with composite strength $T$-scores below 45. This value corresponds approximately to the male 25th percentile ($T = 44.5$) in the sample.

We illustrate this drop-off effect in Table 6. The table shows Pearson correlations and regression coefficients of strength and endurance composites on roof ladder time for subsets of recruits scoring in different strength ranges. The first row of Table 6 summarizes correlations and regressions for the full sample. Rows 2 and 3 provide the same statistics for the 92
male and female recruits scoring below the male 25th percentile for males and for the 209 recruits scoring above the 25th percentile, respectively. Although both are significantly different from zero, the regression slopes of the low- and high-strength groups differ significantly (p < .001) for the strength composite variable. In the low strength subgroup the endurance composite is not significant—strength is the sole performance-based predictor of ladder time for these recruits. The bottom row of Table 6 summarizes regression statistics for recruits in the highest third of the strength distribution. Both strength and endurance regression slopes remain significant and are similar to those for the upper 75% group.

Because of the range restriction on strength, the strength–roof ladder correlations, and consequently the multiple Rs, shrink in the two high strength subgroups, as indicated in Table 6. Because of the nonlinearity related to performance drop-off at low strength levels, the standard correction for restriction of range underestimates the full-sample correlation for these groups. The underestimate is particularly severe (r\text{adjusted} = −.38 vs. r_{\text{full sample}} = −.65) for the recruits in the top third of the strength distribution.

We carried out similar multiple regressions for the remaining criterion variables, using composite strength as a predictor but also including strength group (above or below male 25th percentile) and the composite strength by group interaction term in the equation. A significant slope

### Table 6

<table>
<thead>
<tr>
<th>Sample</th>
<th>Strength composite</th>
<th></th>
<th>Endurance composite</th>
<th></th>
<th></th>
<th></th>
<th>Multiple R</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>r</td>
<td>b ± SE\text{a}</td>
<td>p</td>
<td>r</td>
<td>b ± SE</td>
<td>p</td>
<td></td>
</tr>
<tr>
<td>Full sample (287 males, 19 females)</td>
<td>−.63</td>
<td>−.56 ± .04</td>
<td>.0001</td>
<td>−.36</td>
<td>−.16 ± .04</td>
<td>.0002</td>
<td>.653</td>
</tr>
<tr>
<td>Below male 25th percentile (73 males, 19 females)</td>
<td>−.64</td>
<td>−1.16 ± .14</td>
<td>.0001</td>
<td>−.09</td>
<td>−.04 ± .08</td>
<td>.6606</td>
<td>.641</td>
</tr>
<tr>
<td>Male 25th – 100th percentile (209 males)</td>
<td>−.39\text{b}</td>
<td>−.37 ± .06</td>
<td>.0001</td>
<td>−.33</td>
<td>−.22 ± .05</td>
<td>.0001</td>
<td>.481</td>
</tr>
<tr>
<td>Male 67th–100th percentile (96 males)</td>
<td>−.27\text{c}</td>
<td>−.36 ± .13</td>
<td>.0061</td>
<td>−.43</td>
<td>−.26 ± .06</td>
<td>.0001</td>
<td>.477</td>
</tr>
</tbody>
</table>

\text{a}Raw regression weights ± 1 standard error. All p levels are those associated with the regression coefficients, b.

\text{b}When corrected for restriction of range on the strength composite measure, relative to the full recruit sample, r increases to −.52.

\text{c}When corrected for restriction of range on the strength composite measure, relative to the full recruit sample, r increases to −.38.
difference in high and low strength groups is reflected in the interaction term. Despite the lower power from the smaller samples, results were similar to those found for roof ladder time. For the training rank measure, regression slopes were .019 and .056 for high- and low-strength groups, respectively. Both slopes differed significantly from zero \( (p < .001) \) and differed significantly from each other \( (p = .031) \). For combat test completion time, high- and low-strength group regression coefficients were \(-3.0\) and \(-9.4\) respectively, again both significantly different from zero \( (p < .01) \) and different from each other \( (p = .019) \). For these three criterion measures, the performance drop-off rate was approximately three times greater for recruits falling below the male 25th strength percentile than it was for the higher strength–endurance group. The axe rating did not show the low strength drop-off effect \( (p > .35) \) found for the other criterion measures.

We also looked for blatant false negatives with respect to the instructor choices of their top recruits in each academy class and the 25th percentile strength cutoff. “Top 5 in class” choices were made in five of the nine academy classes, containing a total of 174 recruits. None of the 25 recruits chosen as a “top 5 in class” came from the low strength group, whereas 17 of the 25 rated as poorest in class came from the low strength group. A binary logistic regression of the 25 recruits in the “top 5” group and 25 recruits in the “poorest 5” group indicates that together, composite strength \( (p = .005) \) and height \( (p = .021) \) maximally distinguishes these two groups.

**Results of Structural Equation Modeling**

To examine the validity and usefulness of our field performance assessments of strength and endurance to predict firefighter performance on physically demanding tasks in a more comprehensive manner than simple regression methods permit, we employed FIML structural equation modeling, using Amos 4.0 software (Arbuckle and Wothke, 1999). We used speed in completing the task (i.e., time in sec\(^{-1}\)) to eliminate positive skew found in the time scores. Scores obtained at both mid-training (Ladder 1) and end-of-training (Ladder 2) were used to provide an estimate of test-retest reliability of the measure.

Our initial model was based on the assumption that the underlying physical abilities assessed in our study could be adequately described by two correlated latent variables—strength and endurance. Because the strength and endurance variables were assessed in close temporal proximity by the same instructors, we then tested for improved fit of our model by allowing for correlated error terms among our physical measurements. Modest but significant correlations were observed between grip strength
and other measures and were thus incorporated into our revised model. Figure 2 shows the results of this strength–endurance model used to predict performance of the 287 male recruits on the roof ladder evolution. Values shown in Figure 2 are correlations and standardized regression weights. Setting the correlated error terms, shown at the bottom of the figure, to zero results in only trivial changes in the remaining 12 regression and correlation coefficients, the largest difference being .03.

The model shown in Figure 2 fits the data exceptionally well, with a root mean square error of approximation (RMSEA) of .000 (90% CI: .000–.047), well below the upper reasonable limit of .08 suggested by Browne
and Cudeck (1993). The value of the PCLOSE statistic associated with RMSEA was $p = .97$, which is well above the $p > .50$ recommendation of Jöreskog and Sörbom (1996). Dropping any of the paths shown in the model leads to a significantly poorer fit and attempting to combine strength and endurance into a single latent trait led to a serious loss of fit of the model ($\chi^2_{3 df} = 120.0, p < .0001$).

Together, percent body fat, relative aerobic capacity, and number of sit-ups provide a reasonable assessment of the latent variable of endurance and similarly, strength is largely assessed from the bench press and lat pull measures, with lower but significant loadings on grip strength and number of sit-ups. As indicated, the strength and endurance constructs are moderately correlated ($r = .22, p = .006$). Of primary interest are the regression weights predicting ladder speed, both of which are highly significant ($p < .0001$, using both maximum likelihood estimates (mle) and bootstrapping methods), with strength showing approximately double the weight of endurance (multiple $R = .69$).

Including female recruits in the sample increases variance on both predictor and criterion measures, increasing multiple $R$ to .77. In the full sample of 287 men and 19 women (6.2% women), the regressions of ladder speed on strength and endurance are .64 and .27, respectively. The strength regression coefficient and multiple $R$ continued to increase when we simulated increasingly higher percentages of women in the sample (up to 25%) by over weighting female scores. This pattern of increased regression weights for strength and an increased multiple $R$ when female recruits are added to the sample occurred for all criterion measures where sufficient women were available for inclusion in the analysis. The correlation between gender and endurance was .31 and between gender and strength was .52, reflecting the large gender differences in the scores of the measures that loaded on these two latent traits.

Table 7 gives the standardized regression coefficients for male recruits, using the model shown in Figure 2, for the performance criterion measures used in the study. As would be expected, the relative size of the strength and endurance coefficients differ for the various criterion measures, although strength is highly significant in all cases. The attenuated-adjusted multiple correlations are high for all criterion measures. The lower multiple $R$ for the instructor overall rating reflects, in part, that this measure is not corrected for criterion unreliability.

Finally, we attempted to determine if the three latent performance variables (ladder speed, axe rating, and combat test speed) could be adequately represented by a single latent variable representing all physically demanding fire-ground tasks. We excluded the instructor’s training consensus rank, both because it was not a latent measure and presumably was a judgment
TABLE 7
Latent Trait Standardized Regression Coefficients (Male Recruits)

<table>
<thead>
<tr>
<th>Criterion performance</th>
<th>N</th>
<th>Strength</th>
<th>Endurance</th>
<th>Multiple R</th>
<th>$\chi^2$/df</th>
<th>PCLOSE\textsuperscript{a}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roof ladder speed</td>
<td>287</td>
<td>.58\textsuperscript{b}</td>
<td>.28</td>
<td>.69</td>
<td>0.79</td>
<td>.97</td>
</tr>
<tr>
<td>Axe rating</td>
<td>167</td>
<td>.28</td>
<td>.54</td>
<td>.67</td>
<td>1.41</td>
<td>.66</td>
</tr>
<tr>
<td>Combat test speed</td>
<td>60</td>
<td>.79</td>
<td>.22\textsuperscript{c}</td>
<td>.86</td>
<td>1.51</td>
<td>.59</td>
</tr>
<tr>
<td>Training consensus rank\textsuperscript{d}</td>
<td>167</td>
<td>.44</td>
<td>.07\textsuperscript{d}</td>
<td>.46</td>
<td>1.76</td>
<td>.71</td>
</tr>
</tbody>
</table>

\textsuperscript{a}Closeness of fit of root mean square of error of approximation (RMSEA) of the model. Values of $p > .50$ indicate a good fit (Jöreskog & Sörbom, 1996).

\textsuperscript{b}All regression coefficients and multiple $R$ are significant at the $p < .001$ level unless indicated otherwise.

\textsuperscript{c}$p = .020$.

\textsuperscript{d}Assuming $r_{yy} = .80$ for this criterion measure, $R$ would increase to .52.

Based heavily on performance on the three remaining measures. Because Amos computes full information mle (Anderson, 1957) in the presence of missing data, unbiased and efficient parameter estimates can be obtained despite variation in sample size across criterion measures (Brown, 1994; Arbuckle, 1996). Our model thus incorporated two confirmatory factor analyses—one establishing the constructs of strength and endurance and the second establishing the construct of firefighter performance on physically demanding fire suppression and rescue tasks (physical tasks)—and then related these constructs in a structural model. Figure 3 also shows all five measurement submodels, relating observed and unobserved variables. The paths from the random variation latent variables of predictors are shown in Figure 3 to show the correlations among these variables. For clarity, the remaining random variation latent variables have been omitted from the diagram.

The modeling proved successful (RMSEA = .049, 90% CI: .030–.068; PCLOSE $p = .49$), despite the complexity and scope of the model. As with the previous models using single criterion measures, dropping any of the paths shown in the Figure 3 model leads to a significantly poorer fit and attempting to combine strength and endurance into a single latent trait led to a serious loss of fit of the model. As was the case with the model shown in Figure 2, setting the correlated error terms to zero has a trivial effect on the remaining coefficients.

The loadings of ladder speed, axe rating, and combat test speed on the higher order latent variable are high and approximately equal. All five regression coefficients between the six latent variables are highly significant...
These results indicate that the three rather different criterion measures can be represented by a more general firefighter physical performance latent variable, *physical tasks*, which in turn is substantially predicted by *strength* and *endurance* latent variables ($R = .75, p < .0001$). If we regard the ladder, axe, and combat test evolutions as a representative sample of physically demanding fire suppression and rescue tasks of municipal firefighters, a physical abilities selection test, weighted approximately 60% upper body strength and 40% endurance, based on male score distributions, provides an excellent predictor of success.
Because both the composition of the physical selection tests and the selection ratios of male applicants differed across academy classes, it is not possible to calculate an accurate estimate of range restriction in our S/E assessment, relative to the male applicant pool. Because fewer than 3% of male applicants failed the various physical selection tests, the reduction in recruit S/E SD relative to the male applicant pool is unlikely to be greater than 9.3%. The uncorrected Multiple $R$ increases from .75 to .78 when adjusted for this maximum expected range restriction.

Range restriction adjustments are considerably larger when considering the full mixed-gender sample because the female recruits were highly selected on physical ability. Based on our estimated pass rates of 25% women and 97% men, the original physical selection test $d$ averaged roughly 2.5. Thus, if 20% of the total applicant pool was female, the unrestricted composite strength SD would have been approximately 27% larger than that of our selected sample. Using the model in Figure 3, our mixed-gender regression coefficients of physical tasks on strength and endurance are .64 and .33, respectively, and multiple $R = .81$ ($p < .0001$). Adjusting for our estimated range restriction increases the mixed-gender regression coefficients of physical tasks, on strength and endurance to approximately .69 and .34, respectively, and multiple $R$ to approximately .87, based on an applicant pool with 20% women.

Effects of Cut-off Scores on the Utility of Strength-Endurance Predictors

Because many jurisdictions use pass–fail cut-off points in lieu of top-down selection on physical ability tests, we examined the effect of choosing various cut-scores on utility of a S/E selection procedure for men. The mean $z$-score increase in criterion performance, ($M_{z_{\text{crit}}}$), obtained by a selection procedure, compared to random hiring, is the product of the validity of the selection device ($r_{xy}$) times selectivity, expressed as the mean $z$-score of the selected group on the predictor (i.e., $M_{z_{\text{crit}}} = r_{xy} M_{z_{\text{predi}}}$). When $M_{z_{\text{crit}}}$ is multiplied by some scale of economic value, such as the $SD$ of the dollar value of employee work contributions, one has an estimate of the utility, or practical economic value, of a selection procedure (e.g., Schmidt & Hunter, 1998; Schmidt, Hunter, McKenzie, & Muldrow, 1979). Although even more elaborate extensions of the utility equation are available (e.g., Boudreau, 1991), $M_{z_{\text{crit}}}$ or, equivalently, $r_{xy} M_{z_{\text{pred}}}$ remains directly proportional to utility and is particularly useful when dealing with standardized latent performance variables, such as our physical tasks variable.

Based on the corrected validity coefficient obtained in the structural model depicted in Figure 3 ($r_{xy} = .75$), we can calculate performance gains, in criterion $SD$ units, for various pass–fail cut-off scores, relative to
random hiring. $M_{z-pred}$ is calculated from $\lambda/\varphi$, where $\varphi$ is the proportion selected and $\lambda$ is the ordinate of the normal distribution corresponding to $\varphi$. Selecting recruits in the top third of male applicants would result in a .82 SD performance gain on our latent physical tasks criterion variable, and selecting those scoring in upper half of male distribution would result on a .60 SD performance gain. Low pass–fail cut-off scores substantially reduce utility. For example, selecting recruits with scores in the upper 90% and 95% of the male applicant distribution produces only .14 SD and .08 SD performance gains, respectively, on physical tasks, compared to randomly selected men. Because of the multiplicative relationship between $r_{xy}$ and $M_{z-pred}$, the loss in utility in selecting from the upper 95% of the male distribution, versus the upper 50%, is the equivalent of reducing the validity coefficient from .75 to .10 when selecting from the upper half of the male distribution.

Discussion

Our results indicate that S/E measures assessed using simple, non-laboratory procedures combine to provide substantial predictive validity of firefighter performance on critical job tasks involving high physical effort. Traditional regression analyses show that strength measures are the primary predictors of firefighter task performance, with some incremental validity obtained with the addition of $V_o_{2max}$ and sit-ups variables. These two additional predictors provide slightly more accurate performance distinctions among recruits with relatively high strength levels. Among recruits falling below the male 25th percentile on the strength composite measure, strength was the only significant physical ability predictor of subsequent firefighter task performance for most measures. Strength measures also had higher test–retest reliabilities than our $V_o_{2max}$ assessments. Although relative aerobic capacity contributes only modestly to the incremental prediction of performance on the firefighting tasks studied, its overall importance as a selection variable may be greater when losses due to health problems and injuries are also considered. For example, heart attacks account for approximately 44% of all firefighter fatalities (TriData, 2002), and aerobic exercise appears to reduce the incidence of back and shoulder injuries in firefighters (Beaton et al., 2002). Physiological assessments have generally recommended $V_o_{2max}$ levels of 40–45 ml/kg/minutes for demanding firefighter tasks (Gledhill & Jamnik, 1992; Sothmann, Saupe, Jasenof, & Blaney, 1992).

Rapid execution of many physically demanding fire suppression tasks is essential for job effectiveness in critical situations. Although influenced by many local conditions, described by Thomas (1995), energy output generated from fire spread is related to the cube of time elapsed since
ignition. Recognizing the importance of rapid response, the National Fire Protection Association (NFPA) issued Standard 1710 (National Fire Protection Association, 2004), widely endorsed by firefighting associations, which specifies a response time of 4 minutes for the first arriving fire unit, normally with four firefighters. During the first 3 minutes before the arrival of the full alarm assignment, a typical fire will increase in energy output approximately 400–600%. In addition to escalating property loss, every human risk factor, such as flashover or structure collapse and the probabilities of all forms of injury and death increase, often in an accelerating manner, with time elapsed after ignition. These data show that speed and effectiveness in performing critical fire suppression and rescue tasks is monotonically related to upper body strength and cardiovascular endurance across the full range of $S/E$ scores.

The roof ladder evolution is the primary criterion variable in this study. Data are available from all recruits, and evolutions similar to this are frequently performed, critical firefighting tasks that involve physical demands common to many fire suppression and rescue procedures. Because roof ladder placement is a prerequisite to most roof ventilation procedures, which must precede other fire suppression procedures, time of completion is a performance index directly tied to the magnitude of fire losses. Regression analyses of roof ladder data indicate predictive validity and linearity throughout the upper range of $S/E$ levels, but with evidence of a sharp drop-off in performance for recruits with strength levels below the male 25th percentile. A similar drop-off effect was seen for training rank and the combat test.

These results indicate that top-down selection maximizes the utility of the $S/E$ assessment used in this study. If a pass–fail cutoff were to be adopted, our data suggest that the utility loss of the pass–fail selection strategy will accelerate when the cut-point falls below the male recruit 25th percentile. The greater performance drop-off at the low end of the strength continuum has implications for criterion-related validation studies carried out with incumbents already preselected on strength-related measures. With even moderate preselection, these studies will underestimate the validity and utility of strength indices as performance predictors, even when range restriction corrections are applied, as shown in Table 6.

In our study, approximately 3% of men were screened out, so range-corrected validity coefficients increase only slightly for men. The effects of range restriction were greater for the full mixed-gender sample because approximately 75% of female applicants were screened out by the entry-level physical. Prediction of the latent variable *physical tasks* from the *strength* and *endurance* latent variables increased $R$ from .81 in our restricted sample to .87 after correction for range restriction.
Performance drop-off at low strength levels may be a result of several factors. For many tasks a minimum force is required for effective execution—an axe head traveling below a critical velocity will bounce off a plywood sheet, for example. Progressing downward in a strength continuum, a point may be reached where the force generated by individuals at that level approaches the minimum required to perform a specific task. At that point, time to task completion can increase substantially. In addition, added procedural complexities resulting from compensatory actions for low strength, such as shifting to a two-hand carry or a shift from upper body to lower body muscles, can also be costly with respect to task execution time.

Performance drop-off was not observed for instructors’ ratings axe, which may be an artifact of “central tendency bias,” which appears to be ubiquitous in performance ratings (e.g., Landy & Farr, 1983). Scale compression may be particularly strong at the low end of the performance continuum, as many raters observing exceptionally poor performance are reluctant to assign extreme low ratings and, instead, rank those individuals just slightly below the moderately poor performers. Unlike other measures, the axe performance ratings appear to reflect some age bias. Older recruits received lower power and endurance ratings than predicted from their strength, endurance, and body size.

We found little evidence of high specificity with respect to the strength measure used in prediction. Although the results suggest that one could adopt a single strength measure as a proxy for a more general assessment of upper body strength, the approach might encourage highly focused pretest training. The robustness of multiple-measure assessments of traits compared to best single assessment also argues for the test battery approach. This was demonstrated by the structural modeling results, where the latent strength variable consistently showed higher predictive validity than individual strength measures, even when equated for unreliability of criterion measures. A consistent finding was the modest but significant role of height in predicting criterion performance, even after accounting for strength measures. In the full structural model, however, height did not add significantly to prediction. Identification of measurable skills and abilities associated with height or limb length might provide some small incremental validity to a simple strength-endurance selection test.

Because this study contains only a small, preselected sample of female recruits, we must examine other large sample estimates of the size of gender differences in upper body strength and aerobic capacity to help predict the efficacy of cut-off scores as a selection strategy to reduce the disparate impact of strength measures on female applicants. Because the studies cited below all reported gender differences as female/male ratios, we calculated all $d$ statistics based on the reported means and $SD$s.
Several large-scale, highly standardized, gender difference studies in physical capacities were carried out and reported by the armed services in the 1980s, in response to plans to increase the proportion of women in the military. Together, four of these studies (Knapik, Wright, Kowak, & Vogel, 1980; McDaniel, Skandis, & Madole, 1983; Robertson, 1982; Teves, Wright, & Vogel, 1985) measured upper body strength related to lifts in a total of over 3,000 male and 2,000 female recruits. Standardized mean strength differences, $d_s$, in these studies ranged from 2.8 to 3.8 ($Mdn = 3.1$), with higher lifts (1.9 m) showing the largest differences in weight lifted. Overlap in male and female strength distributions was small. McDaniel et al. provided frequency distributions for several measures, showing that only 2–3% of women exceeded the male fifth percentile and 0–1% exceeded the male 10th percentile. Estimates of overlap in the other studies, based on the reported means and standard deviations (assuming normal distributions), were similar to those in the McDaniel et al. study. Gender differences were smaller for other upper body strength measures, with $d_s$ ranging from 2.4 to 2.6. These military assessments were made before training, but in those studies where both pre- and posttraining strength measures were taken, the magnitude of the gender differences was similar in the two sets of measurements.

Hogan (1991a) summarized gender differences in strength and endurance from five studies in industrial settings that involved a total of approximately 1,965 men and 1,374 women (Table 6, pp. 802–803). Based on tabled $M$ and $SD$, the median gender $d_s$ for upper body strength measures in her table was 2.97, just slightly below the estimate for military populations. Working at the physiological genomic level, Roth et al. (2002) found a substantial gender influence on muscle gene expression, with differential expression greater than 1.7-fold for men, observed for approximately 200 genes.

Reported estimates of the magnitude of gender differences in relative aerobic capacity are less consistent than those found for strength. The inconsistency occurs largely because age, ethnicity, and lifestyle contribute substantially to relative $V_o2_{max}$, hence the degree of heterogeneity in these factors has a large influence on within-gender $SD$ estimates. Thus, in an age homogeneous sample of military recruits (210 men, 212 women, age $SD = 2.1$ yr), the within-gender $SD$ was 4.5 (Vogel, Patton, Mello, & Daniels, 1986). In contrast, a pair of meta-analyses of $V_o2_{max}$ scores in 4,884 women (Fitzgerald, Tanaka, Tran, & Seals, 1997) and 13,918 men (Wilson & Tanaka, 2000) with wide age ranges (age $SD \approx 17.2$ yr) together produced a mean within-gender $SD$ of 9.3. Because we were interested in the gender difference in aerobic endurance, $d_e$, relevant to an applicant population, we estimated the within-gender $SD$ adjusted to match the age
range of the recruits in our study, where 95% were between ages 18 and 35 years.

Using data presented in the two large meta-analyses and a later study by Skinner et al. (2001) with 287 men and 346 women, we estimated the within-gender SD of VO$_{2\max}$ for an 18–35 years applicant pool to be approximately 6.5, slightly larger than the within-gender VO$_{2\max}$ SD of 5.6 obtained for our somewhat range restricted recruit sample. Using the estimated within-gender SD of 6.5, we calculated $d_e$ to be 2.0 for Vogel et al., 1.3 for Skinner et al., and 1.2 based on the two meta-analyses. An additional meta-analysis of age-matched samples (Shvartz & Reibold, 1990) also produced a gender $d_e \approx 1.2$ for 18–30 years olds. Gender $d_e$ for VO$_{2\max}$ thus appears to be approximately 1.2 for a typical firefighter applicant pool.

The results summarized above suggest that for an age-appropriate population, gender $d$ typically falls in the 2.6–3.2 range for upper body strength measures ($d_s$) and is approximately 1.2 for the VO$_{2\max}$ endurance measure ($d_e$). Because strength and endurance jointly contribute to success in performing physical firefighting tasks, most employment tests assess both factors in varying proportions. It is thus useful to have an estimate of the expected standardized gender difference for composite S/E tests. A common misconception is that $d$ for a composite measure is the weighted average $d$ of its individual components, whereas it is a function of the variance–covariance matrix of the component measures. When the test components are in standardized form with unit variances, the $d$ of a composite score based on the weighted sum of two individual components, with weights $w_1$ and $w_2$, can be expressed as:

$$d_{1+2} = (w_1d_1 + w_2d_2)/\left(w_1^2 + w_2^2 + 2r_{12}w_1w_2\right)^{0.5},$$

where $r_{12}$ is the within-gender correlation between components 1 and 2.

We can substitute our estimated lower bound values of $d_s(2.6)$ and $d_e(1.2)$ and $r_{se}(.10$ using VO$_{2\max}$) into equation (2) to calculate $d_{s+e}$ for different weightings of $w_s$ and $w_e$. Equally weighting strength and endurance, for example, results in $d_{s+e} = 2.56$, trivially different from $d_s$ alone.

Although it is a misconception to assume that adding an endurance component to a strength test will automatically reduce $d$ of the composite score, a composite $d$ can be reduced by adding “neutral” physical components that show little or no gender difference ($d_n \approx 0$). For example, in a timed job simulation test, the inclusion of physical activities that make trivial S/E demands but add to variance of time scores will reduce the overall $d$. If a test contained three equally weighted components involving
strength and endurance, plus neutral physical activities uncorrelated with gender, strength, or endurance, equation (1) becomes:

\[ d_{s+t+n} = (w_s d_s + w_e d_e) / (w_s^2 + w_e^2 + w_n^2 + 2r_{se} w_s w_e)^{0.5}, \]  

(2)

thus, the \( d_{s+t+n} \) of 2.56 reduces to \( d_{s+t+n} = 2.12 \). Although the addition of gender neutral physical activities to a test can reduce \( d \), their contribution to test validity depends on their correlation with the criterion. If these are near zero, as Hogan’s (1991a) survey suggests, the reduced \( d \) occurs at the expense of test validity.

The large gender differences described above suggest that using low cut-off scores with tests demanding high strength and endurance is not an effective strategy for reducing disparate impact on female applicants until the cut-off score is set near the male 5th percentile. The point is sometimes made that even an extremely low cut-off score will screen out a few applicants likely to be problematic in subsequent training and job performance, so the strategy has some minimal utility. This argument ignores the likelihood that the small group of failing applicants includes many false negatives, including applicants with a temporary debilitating illness or injury at test time, or applicants disqualified for failure to follow instructions, or other failures unrelated to true physical ability.

Large gender differences create a dilemma for municipalities attempting to set a cut-off score for a physical ability test for firefighters. If the gender \( d \) on such tests is approximately 2.6, as the above results suggest, the male passing rate must exceed 96% for female applicants to attain a 20% passing rate. Thus, in an 85% male applicant pool, approximately 3.5% of the group eligible for further screening will be female. This potential gain in female hires comes at a cost of reducing the utility of physical screening to near zero for men, despite its high validity. Municipalities adopting similar strategies sacrifice most of the test utility in return for very small increase in female hiring. The utility loss stems primarily from the hiring of less physically qualified men than would have been possible with a higher cut-off score or top-down selection.

Because most fire service selection procedures include a test of general intellectual ability, we can compare the outcome of the low physical cut-off score approach just described to a cut-off strategy based on an equally weighted intellectual/physical exam. Assuming no mean gender differences on the intellectual component and a low correlation between components, gender \( d \approx 1.8 \). Setting the composite score cutoff to eliminate the bottom third of our hypothetical 85/15 applicant group would produce a group of eligible candidates that was approximately 3.5% female, as above. An overview and discussion of several issues concerning
multipredictor composites, group differences, and adverse impact can be found in Sackett and Ellingson (1997).

In most critical fire and rescue situations, time to task completion is directly related to the amount of property and human loss, with destruction increasing at an accelerating rate because of the exponential growth in energy output related to fire spread. This unequivocal relationship poses a serious dilemma for researchers attempting to establish a critical score—that specified point in a score distribution at which candidates are considered to have successfully met a minimally acceptable performance (MAP) level for firefighters. Unlike cut-off scores, which “may be established on the basis of a number of considerations (e.g., labor market, organizational constraints, normative information)” (SIOP, 2003, p. 88), critical scores are criterion referenced and empirically determined.

Establishing critical scores is not exceptionally difficult for jobs where the relationship between performance level and job effectiveness begins as a positive monotone but reaches an asymptotic level beyond which further increases in performance do not result in further increases in job effectiveness. But, in jobs where the leveling off of job effectiveness occurs only at high performance levels or when there is no evidence of leveling off throughout the normal performance range, establishing a MAP shifts from being a criterion-based critical score determination to a broader and more problematic cost/benefit-based, cut-off score determination.

In the stark terms of this fire service example, experts would first have to quantify the relationship between performance level on critical tasks and the resulting life and property loss, and then senior management would determine how much loss is tolerable when setting the MAP standard. Experts would likely find the loss assessment step extraordinarily difficult and rarely technically feasible, and senior management would likely regard making a no-win MAP decision onerous, even if an adequate study were available.

Establishing a MAP level for applicants for safety forces positions re-emerged with some salience as a result of the two Lanning v. Southeastern Pa. Transportation Authority decisions in the United States Court of Appeals for the Third Circuit (1999, 2002). In the Lanning I decision (1999) the court raised the standard for analyzing the business necessity defense by requiring that the employer demonstrate that only the minimum qualifications necessary for the successful performance of the job are being assessed:

A business necessity standard that wholly defers to an employer’s judgment as to what is desirable in an employee therefore is completely inadequate in combating covert discrimination based upon societal prejudices. Only a business necessity doctrine that examines discriminatory cut-off scores
in light of the minimum qualifications that are necessary to perform the job in question successfully can address adequately this subtle form of discrimination. (p. 15).

Using the SIOP 2003 definitions, a narrow interpretation of the strict standard outlined in Lanning I (1999) would suggest that only criterion-based critical scores would be acceptable evidence of business necessity, whereas most physical ability MAP decisions involved in fire service hiring are cut-off score decisions involving substantial employer judgment. If a positive monotonic relationship exists between performance level and job effectiveness in any job, MAP decisions necessarily become largely employer judgment decisions deemed unacceptable by the Lanning I Court. In Lanning II (2002) however, the Circuit Court suggested that the minimum qualifications standard may not be as strict as that implied in Lanning I (e.g., see Sarno, 2003). The discrepancy between the Circuit Court’s strict theoretical standard in Lanning I and its more lenient practical standard in Lanning II is troublesome for those involved in safety forces selection. Employers and test developers may conclude that the theoretical standard is impossible to meet and therefore choose not to risk adopting certain high utility selection procedures.

A simple alternative to the difficult procedure of relating physical ability to cost-estimated losses and using the regression to establish a MAP is to have subject matter experts (SMEs) observe a range of performance levels and choose those they find acceptable. Unlike an objective criterion obtained from a job analysis, such as the maximum weight required to be lifted on the job, a performance criterion derived from SME judgments is likely to be based on many of the same considerations involved in establishing cut-off scores. In some cases, these cost/benefit decisions are inappropriately relegated to job incumbent SMEs rather than senior management. In addition, because career-length fitness testing of incumbents has long been an issue for safety forces, using incumbent judges to subjectively establish a MAP standard may create a sizeable conflict of interest. Even when judges are appropriately chosen, MAP rating procedure methodology is fraught with difficulties and often unfounded assumptions.

Some of these problems are exemplified in a study by Sothmann et al. (2004), who claimed to have “validated minimally acceptable standards for the muscular strength and endurance necessary to perform fire suppression activities.” Incumbents viewed four video recordings of firefighters performing a series of evolutions at different paces and indicated whether each pace shown was “acceptable” or not. All four task completion times in the videos were longer than the median time of an age-heterogeneous group of male and female incumbent firefighters with little incentive to perform near their maximum capacity. Given a substantial literature showing
the influence of stimulus conditions on judgments (e.g., Anderson, 1975; Murphy & Costans, 1987; Parducci, 1965, 1990; Poulton, 1979) it is likely that the MAP obtained from SME ratings of these slow-pace video stimuli is biased towards poor performance. In addition to being influenced by anchor and stimulus characteristics, rater judgments are often context dependent, yet no context was provided. It is unlikely that the same pace level would be chosen if SME instructions indicated that the fire involved an abandoned structure versus instructions describing a residence fire with victims trapped on the second story.

The video rating method also depends on informed raters to provide judgments that are reasonably consistent and bias-free. The Sothmann et al. study involved oversampling of women and older male incumbents as judges, chosen after many of their co-workers had been tested and received time scores on the recorded exercise, yet there is no reference to any examination of judge behavior. Finally, MAP was arbitrarily defined to be the slowest paced video that 50% of the judges found acceptable. A slightly higher standard of agreement or a change in the decisions of just two judges would have reduced the MAP time from 485 to 426 seconds, resulting in a substantial drop in the number of incumbents who would have been considered acceptable by their coworker judges. The authors’ choice of a 485 sec MAP thus appears to have a rather fragile basis, especially when 30% of the judges failed to correctly rank order the pace shown in the four videos. Study methodology appears to have been a major determinant of the MAP chosen by the authors. Although researchers attempting to set MAP levels can avoid many of the shortcomings of the Sothmann et al. approach, issues such as choice of SME judges and defining “consensus judgment” will still require resolution.

There is one additional issue involving minimum hiring standards that is specific to physical ability testing. Implicit in establishing a MAP level for job applicants is the assumption that physical performance will not decline over the expected period of employment, particularly for candidates scoring close to the MAP cutoff at the time of hire. Without compulsory or high incentive employee fitness programs and career-length MAP retesting, this assumption is untenable. Active men and women respectively show average declines in \(V_o_{2\text{max}}\) of approximately 4.0 ml/[kg minute] and 4.4 ml/[kg minute] per decade, along with increased body weight (Fitzgerald et al., 1997; Wilson & Tanaka, 2000). Although absolute strength declines are modest until age 50 (e.g., Lindle et al., 1997; Metter, Conwit, Tobin, & Fozard, 1997), age-related weight gain during this period results in declining relative strength.

Results obtained from our structural equation modeling indicated that there is a great deal of generality in both predictor and criterion measures for firefighting activities. A general firefighting physical task performance
factor adequately accounts for most of the reliable variance in the different criterion measures, and a two-factor, strength–endurance measurement model accounts for a large proportion of the variance among the predictor variables. Together, strength and endurance predict physical task performance quite well, with a multiple $R$ of .75 for trained male recruits about to enter fire service. Using the full mixed-gender sample in the model increases $R$ to approximately .87, when adjusted for range restriction based on the applicant pool. Among the specific criterion variables, strength and endurance produced reliability-adjusted validity coefficients ranging .50–.85 for the various individual performance measures for the male recruits.

A commonly raised issue in safety forces litigation centers on screening for physical ability levels necessary to perform highly critical but low frequency tasks. Our results indicate that for firefighting the issue may be moot. A single factor accounts for a substantial proportion of the variation in performance in many different physically demanding fire suppression and rescue procedures. In turn, strength and endurance are excellent predictors of that latent physical task variable. Although some of the individual tasks in the firefighting and rescue evolutions of the type studied here may be required infrequently in some municipalities, taken together as a related class of activities, they are performed frequently in most large fire departments.

The existence of a common physical task latent variable that accounts for more than half of the reliable variance in performance on physically demanding procedures also has implications for transportability of validity evidence for fire service. A prerequisite for transportability of a valid selection procedure to a new setting is to show “job comparability in terms of content or requirements” (SIOP, 2003, p. 27) or that incumbents “perform substantially the same major work behaviors, as shown by appropriate job analyses” (EEOC, 1978, p. 206). To demonstrate job comparability, practitioners often look for high correlations across the two employment settings for task frequency ratings and task importance ratings obtained from the two job analyses. There is, however, no consensus on how high such correlations must be to meet the intent of the Uniform Guidelines or SIOP Principles. Discrepancies between municipalities’ ratings of the importance/frequency of various physically demanding fire suppression procedures are far less consequential when the underlying abilities required for effective performance on those procedures are largely the same.

Results of the structural equation modeling provided strong support for the construct validity of a strength–endurance model for predicting performance on a wide range of fire suppression and rescue tasks in trained recruits. This finding is consistent with the results of Davis, Dotson, and Santa Maria (1982), who correlated many S/E variables with time measures of five sequentially performed firefighting tasks; and with the
endurance–performance correlations reported by Rhea, Alvar, and Gray (2004); and with the Sothmann data cited above.

Our analysis also indicated that a general physical performance factor accounts for most of the reliable variance in the separate firefighting tasks assessed at the fire academy. This finding contradicts the “specificity” arguments made to challenge the validity of physical ability selection tests. Such arguments claim that neither predictor measures nor criterion measures are highly correlated, therefore, small variations in the choice of measures can have large effects on predictive validity. Data supporting these arguments often come from highly selected and atypical populations, such as elite athletes, where restriction of range is extreme. We find no such specificity in our more typical sample of firefighter applicants. Our results suggest that one has reasonably wide latitude in choice of measures that will adequately reflect the three underlying constructs of strength, endurance, and firefighter physical tasks performance. We also found that beginning-of-training and end-of-training physical assessments produced nearly identical results when used as predictors of performance, providing no evidence of either decay in the validity of predictors over time or of ability × training interaction effects that would compromise validity coefficients.

In a meta-analysis of written firefighter selection tests based largely on general intelligence and mechanical comprehension, Barrett, Polomsky, and McDaniel (1999) reported an estimated validity coefficient of .77 for prediction of training success and .56 for prediction of subsequent rated job performance. Cognitive tests thus account for approximately a third of the variance in firefighter performance, presumably that related largely to the domain of job knowledge and fireground decision making. Our results suggest that an assessment of $S/E$ can be equally effective in predicting firefighter success in tasks dependent on high levels of physical ability. Because cognitive and physical test scores are usually uncorrelated, it would appear that corrected validity coefficients of equally weighted cognitive and strength–endurance physical tests could be expected to exceed .70. Because selection ratios for firefighting positions in large cities are often below 20%, the potential utility of a composite cognitive–physical test used in top-down hiring is exceptionally high. This extraordinary utility is largely lost with low cut-off scores, especially in two-step pass–fail systems with high pass rates at each step.

**Practical Implications**

Our results raise several practical issues regarding the validation and use of physical ability testing in general and specifically with respect to firefighter selection. First, despite the considerable administrative
advantages, high reliabilities and substantial validities of strength and endurance measures reported here and in Hogan (1991a), it is unlikely that many hiring agencies, especially those involved with safety forces, will abandon job simulation tests soon, despite their shortcomings. Because simulations are generally perceived to be both more job-related and fair by applicants and the courts than strength and endurance measures (e.g., Hogan & Quigley, 1986; Ryan, Greguras, & Ployhart, 1996), substantial validity evidence for S/E tests will be required before these measures begin to replace or supplement currently used simulation tests.

We suggest that the approach used in our study, wherein S/E assessments are made at job entry and then used as predictors of training and job success, may be a simple and effective way to accumulate such evidence. Both isometric strength tests (e.g., Blakley et al., 1994) and isotonic 1-RM strength tests, recommended by The American Society of Exercise Physiologists (Brown & Weir, 2001), provide highly reliable measurements in units that are directly comparable across studies. Similarly, field measures of aerobic endurance, such as the step-test (e.g., Cotten, 1971; Davis & Wilmore, 1990) or 1.5-mile run (e.g., Larsen et al., 2002) are readily converted into estimates of $V_{O_2 \text{max}}$ in liters per minute or relative $V_{O_2 \text{max}}$ in l/[minute kg]. Correlations of S/E measures with the simulation test scores used to hire the new employees can provide insight about the S/E demands of the simulation and can be used in conjunction with the applicant pool simulation $SD$ to provide range restriction estimates of S/E in the selected group (Sackett & Yang, 2000). In many cases, these highly standardized measures may be able to be applied directly to population norms for range restriction corrections (e.g., Hoffman, 1995).

In this regard, Tukey’s (1969, p. 80) comment is germane, with regard to predictors—“Measuring the right things on a communicable scale lets us stockpile information about amounts. Such information can be useful, whether or not the chosen scale is an interval scale.” Stockpiling of information will also require improved reporting of results relating physical abilities to job effectiveness. Too often means, $SD$s, and correlations are omitted or only partially reported, replaced by multivariate summary statistics or verbal generalizations. Correlations are also frequently based on mixed-gender samples that vary in female–male proportions. Although informative for that specific study population, the information is less useful for data stockpiling and providing cross-study comparisons.

Second, even though the latent variables of strength and endurance continue to emerge as reliable constructs with substantial predictive power for specific job tasks, there is much less information available with respect to criterion performance constructs in the physical domain. We found strong evidence for a general factor reflecting physical performance on three physically demanding, extended firefighting/rescue evolutions that
were composed of different mixes of specific job tasks. Because all three evolutions were composites of several tasks with varying strength and endurance demands, a single performance component emerged. Had a set of individual job tasks rather than multitask evolutions been used as criterion variables, a pair of correlated latent performance variables would likely have emerged, reflecting underlying strength and cardiovascular fitness demands of the individual tasks. Arvey et al. (1992) used a construct validation approach using supervisor ratings of performance on specific job tasks for police officers and obtained this result. For the most part, however, work on latent structure of physical criterion variables, especially objectively measured work samples, remains sparse despite the value of such information for issues of transportability of physical tests.

Third, gender differences in physical simulation tests that make high demands on strength and endurance are so large that the two goals of achieving gender representation and maximizing test utility are substantially at odds with each other. In performance tests not infused with variance-increasing “noise” from nondemanding physical activities, female and male applicant groups can be expected to differ by more than two standard deviations. With such large gender differences, compromise strategies such as setting low physical cut-off scores often result in trivial increases in gender representation at a high cost in selection utility. In the face of the immutable connection between time-elapsed and increased property/life loss due to fire, attempts to establish an acceptable minimum level of physical performance must necessarily be trade-off decisions, often dependent on local circumstances. For example, because fire departments differ in the number of firefighters assigned to an apparatus (typically 3–5), officers in different municipalities have varying latitude in their ability to match assignments to firefighter abilities. Minimal staffing of each fire apparatus and increased response times resulting from a more limited network of fire stations relative to workload are more often found in less affluent municipalities. Consequently, these communities are the most likely to be adversely affected by low cut-off hiring decisions.

Fourth, the accumulated evidence associating high strength and endurance with more effective performance in physically demanding critical fire suppression and rescue procedures has implications for personnel policies within municipal fire departments. High-incentive training programs to help incumbents maintain strength and endurance are not only likely to be cost effective in terms of job performance and reduction of lost time related to injuries and coronary artery disease, they may be essential to justify a minimum acceptable performance hiring standard.

Finally, from the research standpoint, this study demonstrates that by incorporating FIML methods in the analysis, one can extract valuable information from employment data containing large blocks of missing
information. Because the methodology is especially powerful for longitudinal and multigroup modeling with missing data (e.g., Little & Rubin, 1989; Wothke, 2000), it is ideally suited for predictive validation studies using job incumbents. Used in conjunction with structural equation models, both criterion-oriented and construct-oriented validation strategies may be possible in situations where traditional regression and factor analytic methods would render a study infeasible because of both low power and bias resulting from missing data.

REFERENCES


Shvartz E, Reibold RC. (1990). Aerobic fitness norms for males and females aged 6 to 75 years: A review. Aviation, Space and Environmental Medicine, 61, 3–11.


