JOURNAL OF APPLIED PSYCHOLOGY MONOGRAPH

Development of Physical Ability Tests for Police Officers: A Construct Validation Approach

Richard D. Arvey and Timothy E. Landon
Industrial Relations Center, University of Minnesota

Steven M. Nutting
Minneapolis Civil Service, Minneapolis, Minnesota

Scott E. Maxwell
University of Notre Dame

A construct validation approach was followed to affirm that 8 physical ability test events were significantly related to two important constructs underlying the job performance of police officers: strength and endurance. A sample of 115 incumbent police officers took 8 physical ability tests and were rated by supervisors on their physical performances in their job. LISREL methods were used to test the model specified, and a reasonable fit was achieved. Portions of the model were tested on an independent sample of 161 applicants; the fit of the model was again acceptable. A nomological network of relationships, in which strength and endurance factors correlated in expected directions with other physiological and demographic variables, was hypothesized and tested. Finally, the data were examined for potential gender differences and bias. Considerable differences were shown between men and women on both test and performance variables, and women would be overpredicted if a common regression line were used for selection purposes.

A contemporary issue in the domain of staffing and selection involves the use of physical ability tests for selecting employees into physically demanding jobs, such as police officer and firefighter. Fleishman (1988), for example, commented that although physical performance tests have existed for a considerable time, it is only recently that such tests have actually come into use in the employment situation. Another cue that the role of physical ability testing in employment setting is of increasing importance is the appearance of a chapter on physical abilities in a recent volume of the second edition of the *Handbook of Industrial and Organizational Psychology* (Hogan, 1991a).

However, the use of such testing procedures introduces potential equal employment liabilities. Both Hogan (1991a) and Campion (1983) have noted that such tests are likely to adversely affect women and that organizations may be asked to defend the validity of these testing procedures. Indeed, a substantial number of legal cases have dealt directly with the use and impact of such physical ability tests on protected group members (Hogan & Quigley, 1986). In a similar vein, the use of physical ability tests will most surely be closely scrutinized within the context of the recently enacted Americans With Disabilities Act (1990), which requires employers to pay attention to the molecular physical requirements of jobs.

Given this increase in importance and possible litigation, what kind of evidence exists for the validity of physical ability tests in employment settings? Hogan (1991a) reviewed 14 studies in which correlations between various physical ability tests and some kind of criterion measure (i.e., training time, job tenure, ratings of job performance, and work sample performance) were reported. The correlations were generally significant and high, especially when the criterion measure involved some work sample measure. In her review, Hogan described an unpublished study by Lewis (1989), who used meta-analytic procedures to compute corrected mean validities for physical ability tests across 13 of these 14 studies. Categorizing the physical ability tests as metrics of muscular strength, anthropometric variables, and muscular endurance, respectively, Lewis's results indicated estimated corrected mean validities of .23, .23, and .30 against training criteria, and .82, .49, and .37 against work sample criteria. Thus, the limited literature supports an empirical linkage between broad types of physical ability tests and a variety of employment criteria. However, there simply is not a great deal of research establishing these relationships. Both Campion (1983) and Hogan (1991a) called for additional research to verify these linkages.

A common strategy used by organizations to establish this

Support for this article was provided by the Center for Urban and Regional Affairs.

We wish to thank Colleen Barberio for her help on this project and Cheri Ostroff, Paul Sackett, John Moore, and Ray Noe for comments on an earlier version of the article.

Correspondence concerning this article should be addressed to Richard D. Arvey, Industrial Relations Center, University of Minnesota, Minneapolis, Minnesota 55455.

Journal of Applied Psychology, 1992, Vol. 77, No. 6, 996-1009
 0021-9010/92/$3.00

linkage is to adopt a content validation approach. It appears, however, that when evidence of this nature is presented, a number of problematic issues and concerns surface. Hogan (1991a), Arvey, Nutting, and Landon (1992), and Bernardin (1988) outlined some of these problems, including the following: (a) assertions that job analysis fails to adequately tap relevant physical duties and performance requirements, (b) assertions that physical ability tests over- or underrepresent relevant aspects of the job, and (c) concerns over the lack of fidelity between the response aspects of the test and the actual response requirements of the job. These and other issues suggest that it can be difficult (and expensive) to defend the use of physical ability tests on the basis of traditional content validation principles and procedures.

In light of this background, it is reasonable to consider the use of construct validation in developing evidence to demonstrate that inferences based on scores derived from physical ability tests are indeed job related. The purpose of the present article is to illustrate the application of the construct validity paradigm to the validation of physical ability tests in a particular applied setting—the selection of entry-level police officers in Minneapolis, Minnesota. We believe this effort is important for the following reasons: First, there are few previous efforts to validate physical ability tests with a construct validity approach even though the nature of the constructs being measured lend themselves well to this strategy. Second, reported formal applications of a construct validity approach to the validation of selection tests in general are relatively rare (see Schmitt, 1988; Vance, Coovert, MacCallum, & Hedge, 1989). Third, there are certain advantages inherent in using a construct validity approach in that this strategy relies on multiple sources of information, utilizes multiple methods, and, perhaps most important, contributes to greater scientific understanding of the relationships observed. Thus, we believe that organizations faced with defending the validity of physical ability tests might profit by using this approach rather than relying solely on either a content or criterion validation strategy.

Another objective in the present research effort was to examine the set of physical ability tests with regard to whether they might be relatively unfair or biased against women applying for a position as an entry-level police officer.

Method and Results

The study was conducted as part of an effort to develop and validate a set of physical ability tests to help select entry-level police officers. The study was conducted in several steps.

Step 1: Specifying the Physical Components of Job Performance

The first step of our process was to develop a rough idea of the kinds of physical activities associated with the entry-level police job. Our intent here was to learn about the various physical activities performed and their relative importance and frequency. However, these data were not intended to be the sole basis for predictor development and defense. Because of the limitations of job analysis methods and data (noted earlier), the information we gathered in this first phase was to provide guidance concerning the kinds of predictor tests to be used as well as to help develop hypotheses concerning the kinds of underlying constructs involved.

Job analytic data were gathered with three different methods. First, the department maintained records, called *Use of Force Reports,* that described any occasion in which officers used physical force on the job. Three hundred thirty-two such records from 1988 were available for content analysis. Two individuals (Richard D. Arvey and a research assistant) worked together to content analyze these reports. They first read about 100 of the incidents and formed 15 preliminary categories of physical activity (e.g., wrestling, pushing, pulling, running, etc.). Subsequently, the research assistant sorted the Use of Force Reports into these categories, and Richard D. Arvey confirmed or changed judgments when necessary. When an incident involved two or more physical activities (e.g., both wrestling and running), it was classified into what seemed to be the most difficult or effort-driven activity. Several categories were collapsed because of overlapping activities. Table 1 presents the number of Use of Force Reports that were classified into the 12 categories that included the greatest number of those reports. The physical activities most frequently represented were wrestling ($n = 126$), pushing ($n = 60$), pulling ($n = 42$), and running ($n = 34$).

Second, 12 experienced officers were asked to generate detailed descriptions of actual job events in which they themselves had used physical effort. These officers generated 45 incidents, which were examined and classified into the following content areas: running and chasing suspects, wrestling and subduing suspects, jumping and crawling, lifting or dragging, kicking, and working in extremely uncomfortable weather conditions. These data were helpful in permitting the inference that such physical activities did indeed occur on this job.

The third method of data collection was a survey instrument, which was developed and sent to a random sample of 50 officers. The survey consisted of 72 items of physical activity. For example, the first item read, "Carry a person who has been arrested and is unable or refuses to walk by grasping the person under the arms, supporting his weight, and transporting him to the police car." The items were developed from the incidents described in the Use of Force Reports and the critical incidents described by the 12 police officers and were organized around the 12 broad physical activity domains identified earlier (e.g. carrying, running, lifting, wrestling, etc.). Officers were asked to rate the frequency and importance of each task. Responses were received from 20 officers. The mean age of this sample was 36.61 years ($SD = 9.48$), which did not differ significantly from the larger population of officers ($N = 448$, $M = 36.35$, $SD = 9.53$) in the department. Table 2 presents the 12 tasks rated as most important (1 = *slightly important,* 5 = *critically important*) and the 14 tasks rated as performed most frequently (1 = *rarely,* 5 = *continuously*).

To summarize, the information gathered revealed a number of physi-

Table 1
Content Analysis of 1988 Police Department Use of Force Reports

Category	Number of reports
Carrying	3
Running	34
Lifting	13
Climbing	7
Pulling	42
Dragging	5
Jumping	3
Wrestling	126
Pushing	60
Kicking	19
Striking	29
Frisking	31

Table 2
Top 12 Job Analysis Items in Terms of Rated Importance and Frequency

Item	*M*	*SD*
Importance		
1. Fire a shotgun at a person or animal.	4.79	0.68
2. Fire a weapon at a person or animal.	4.79	0.68
3. Use sufficient grip strength to fire a gun.	4.79	0.68
4. Carry an equipment belt (which contains a gun, mace, and a walkie-talkie) for the duration of the shift.	4.79	0.51
5. Pull the slide on an automatic handgun or a shotgun to chamber a round.	4.72	1.27
6. Jump out of the way of a car by moving quickly to one side or the other.	4.53	0.80
7. Jump out of the way of an object that has been thrown at you, by moving quickly to one side or the other.	4.37	0.85
8. Administer cardiopulmonary rescuscitation to an unconscious person.	4.37	1.27
9. Use a baton or other device to subdue an attacking or resisting person.	4.26	0.83
10. Wrestle, fight, and subdue attacking or resisting persons using locks, grips, or holds (not including mechanical devices).	4.11	0.89
11. Physically restrain suspect by applying holds, for example, arm bars, neck and shoulder holds, and so fourth (not including mechanical devices).	4.05	0.92
12. Use sufficient grip strength to hold a resisting person.	4.05	0.97
Frequency		
1. Carry an equipment belt (which contains a gun, mace, and a walkie-talkie) for the duration of the shift.	4.95	0.22
2. Pull the slide on an automatic handgun or a shotgun to chamber a round.	3.79	1.21
Frequency (*continued*)		
3. Pull one person away from another to stop them from fighting, by grasping the person and moving him or her away from the other person.	3.68	0.95
4. Carry a person who is unconscious, drunk, or overcome by smoke, by grasping the person under the arms, supporting his or her weight, and transporting him or her to the police car.	3.63	0.91
5. Carry a person who has been arrested and is unable or refuses to walk by grasping the person under the arms, supporting his or her weight, and transporting him or her to the police car.	3.58	0.99
6. Use sufficient grip strength to hold a resisting person.	3.55	1.07
7. Wrestle, fight, and subdue attacking or resisting persons using locks, grips, or holds (not including mechanical devices).	3.37	1.06
8. Run after a suspect on the street to apprehend and arrest a suspect.	3.37	0.96
9. Pull a suspect who is resisting away from an object he or she is holding onto (a car door handle, etc.).	3.37	1.16
10. Physically restrain suspect by applying various holds, for example, arm bars, neck and shoulder holds, and so forth (not including mechanical devices.	3.37	1.01
11. Drag an unconscious person who is drunk.	3.26	1.09
12. Lift a person into the back of a squad or van who is resisting.	3.26	0.83
13. Run after a suspect through the yards of a residential area to apprehend and arrest the suspect—avoiding clothes lines and other obstacles in yards.	3.21	0.93
14. Run up or down stairs to reach the scene of an emergency—or while chasing a fleeing suspect.	3.16	0.91

Note. For importance ratings, 5 = *critically important,* 4 = *very important,* 3 = *important,* 2 = *moderately important,* 1 = *slightly important,* and 0 = *not important.* For frequency ratings, 5 = *continuously, daily;* 4 = *frequently, weekly;* 3 = *occasionally, monthly;* 2 = *seldom;* 1 = *rarely;* and 0 = *not applicable, not performed.*

cal activities that were performed or were important in police work. We also collected job analysis information from other jurisdictions (e.g., the State of New Jersey and the city of Des Moines, Iowa), as well as information concerning the kinds of physical ability testing procedures used and any supporting technical information. Our data appear to be in line with several other job analytic studies conducted on police or state troopers (Booth & Hornick, 1982; Jeanneret, Moore, Blakey, Koelzer, & Menkes, 1991). However, our job analysis was a relatively low-resource effort. Blanket sampling methods were intentionally not used, time-intensive observation methods were not employed, and so forth. Our effort here was simply to gather information concerning the domain of physical tasks that are performed and deemed important in this job and not to provide detailed molecular information concerning tasks and the level of effort expended.

Step 2: Specification of a Theory of Physical Abilities That Underlie Job and Test Performance

Another step in the process was to develop our theory of physical ability constructs as related to the physical components of the job. To this end, we reviewed the current literature concerning the taxonomic structure of physical abilities in general as well as those specifically related to this job. We first noted that literature involving police settings suggests that two physical constructs—strength and endurance—are primary factors underlying job performance and the avoidance of injuries (Hoover, 1989; Maher, 1984).

Second, we noted that Fleishman (1964) developed a taxonomy that has served as the benchmark for physical ability measurement and understanding for many years. He identified nine physical factors and established test events to serve as measures for each factor. However, until recently little work was done to discover any underlying relationships, or structure, among these nine factors. Recently Hogan (1991b) reported the results of factor-analytic work showing that three primary factors or constructs can account for a considerable amount of variance in the performance of physical job tasks and selection tests and in the ratings of physical requirements for job tasks. She interpreted these primary factors as strength, endurance, and movement quality (which included coordination, balance, and flexibility).

On the basis of this literature, we theorized that two of the underlying physical constructs accounting for variance in the job performance of police officers and in their performance on physical ability tests are strength and endurance. For the purposes of this study, we identified strength as the ability to exert physical force against a load (i.e., mov-

ing, carrying, or propelling one's own body or some external load or both). Strength can be thought of as an individual's capacity to exert force or effort at a single point in time. As defined here, strength encompasses Fleishman's (1964) static strength and explosive strength factors. We identified endurance as the ability to sustain or recover from the exertion of effort over time. This includes Fleishman's dynamic strength factor (the ability to exert muscular force repeatedly or continuously over time) and stamina factor (sustained physical effort involving the cardiovascular system). Our use of endurance thus involves the ability to use both (a) short-term anaerobic energy-conversion processes to sustain effort for tasks lasting from a few seconds to about 2 min and (b) long-term aerobic energy-conversion processes to sustain effort for longer than 2 min.[1]

Our view then was that the physical tasks involved in entry-level police work and the physical ability tests we developed could be arrayed on these two dimensions—strength and endurance. Tasks and tests that require at least a moderate amount of instantaneous effort were expected to have a significant strength loading; tasks and tests that require effort sustained more than just a few seconds were expected to have an endurance loading. Hence, we hypothesized that many of the physical job tasks and physical ability tests would load on both strength and endurance. We did not deal with Fleishman's (1964) other physical factors (those involving balance, coordination, or flexibility) nor did we deal with Hogan's (1991b) third primary factor (movement quality). We felt that the influence of these factors on variance in the performance of police tasks would be small and less important compared with the influence of strength and endurance. Moreover, we would have needed to add several additional test events and performance rating items to the information net gathered.

Step 3: Manifest Variable Specification

A variety of information and data were gathered on a specified sample of officers. These data included scores on a battery of physical ability selection tests, criteria ratings of each officer's performance on a variety of physical activities, and data from archival sources, including a file rating of physical performance, as well as data based on annual physical assessments of the officers by the local YMCA.

Physical ability test events Our choice of physical ability tests was guided by both content- and construct-oriented considerations as well as our theoretical framework. Several test events were developed as rough representations of tasks indicated as important in the job analysis (i.e., the dummy drag, dummy wrestle, and obstacle course events). In developing these events, we tried to strike a balance between fidelity and practicality. Several other tests were included because they have served as traditional measures of the abilities thought to underlie test and job performance (i.e., the 100-yard dash, grip strength, situps, bench dips, and a 1-mile run). These are commonly used and well-researched physical tests, and although they are not by themselves isomorphic with physical job tasks, they can be linked to the physical constructs that we hypothesized to underlie performance on the physical tasks involved in police work.

The following tests were included in the selection battery:

1. 100-yard dash (timed in seconds). Job analysis information indicated that running short distances at full speed is an important and frequent component of job performance. This test was hypothesized to have moderate loadings on both strength and endurance.

2. dummy drag (timed in seconds). This test requires pulling or carrying a 120-lb (54.4 kg) dummy a distance of 50 ft (15.24 m). This is a content-oriented test and was hypothesized to have a high strength loading and a moderate endurance loading.

3. obstacle course (timed in seconds). This test involves jumping a hurdle and a simulated ditch, zig-zagging through some markers, crawling under an obstacle, and climbing over a 6-ft (2.09 m) fence while covering a distance of about 60 yards (54.86 m). This is a content-oriented test and was hypothesized to have a moderate strength loading and a high endurance loading.

4. grip strength (measured in pounds of force). This measure is taken with the dominant hand using a dynamometer. Several job tasks required hand and grip strength. Such tests are commonly used in personnel selection and have been shown to be highly related to other tests of maximal strength while posing less risk of injury to the participant. It was hypothesized to have a high strength loading and no endurance loading.

5. dummy wrestle (timed in seconds). This test requires participants to pick up an 80-lb (36.29 kg) dummy and make a series of moves with it (i.e., rotate it, roll with it, and place it on a designated spot). This is intended to simulate contact and wrestling with suspects, and it was hypothesized to load heavily on strength and moderately on endurance.

6. sit-ups (the number of repetitions performed in 1 min). This test is a classic measure of dynamic trunk strength and was hypothesized to have a high endurance loading, with no strength loading.

7. bench dips (number of repetitions performed in 1 min). This test requires participants to face the ceiling, place their feet on one bench and their hands behind and beneath them on another bench, and lower and raise themselves using their arms only. This event was hypothesized to be an upper-body dynamic strength measure, with high endurance but no strength loadings.

8. 1-mile run (timed in seconds). This test was hypothesized to have a high endurance loading as a classic test of cardiovascular stamina as well as dynamic leg strength.

Although the 100-yard dash, the dummy drag, and the dummy wrestling events have a relatively close linkage to the job analysis results, the other events are less tied to police work. Though included because of their theoretical linkage to the central constructs, they should be considered experimental in nature.

The test events were administered in an indoor arena during a one-day period. Each subject was assigned a monitor, who instructed him or her on each test and recorded his or her performance. The tests were administered in the order listed. Time was provided between tests for the officers to rest before starting the next event.

Ratings of physical performance on the job. The immediate supervisor for each of the 200 officers in our sample was contacted and asked to complete a performance evaluation form consisting of scales for rating physical job activities. These items were developed from the job analysis; superiors were asked to rate actual job performance on the following tasks: running, wrestling, lifting and carrying, climbing, crawling and balancing, and pushing and pulling. Other items asked for ratings of endurance, general physical fitness, and overall job performance (which included both the physical and nonphysical components of the job). On the basis of theory and the job analysis information, wrestling, lifting and carrying, and pushing and pulling were hypothesized to be associated with a strength factor, and general physical fitness, climbing, crawling, and endurance were expected to be associated with an endurance factor. Ratings were made on a 1 to 5 scale ranging from *poor performance* (1) to *superior* (5); each scale was verbally anchored with examples of behavior that exemplified the various levels of performance. A single item asking the raters how confident they were about the ratings was also included (rated on a 1 to

[1] Fleishman (1964) did not include an anaerobic factor in his taxonomy. However, research from exercise physiology suggests that although specific physical training activities can be targeted to improve either anaerobic fitness or aerobic fitness, for most individuals, the correlation between these two energy sources is sufficient to warrant combining them in a common endurance factor (cf. McArdle, Katch, & Katch, 1981).

3 scale). The mean value for this variable was 2.31 ($SD = 0.79$), indicating that, for the most part, supervisors felt confident in the ratings they made.

Physiological data. The department had contracted with a local YMCA to conduct annual physical fitness assessments for officers. These assessments included a measure of maximum oxygen uptake capacity, which was made using a treadmill exercise with graded increases in workload to the point of exhaustion. This measure is considered to be a benchmark indicator of aerobic stamina. Also, several anthropometric measures were taken, including height, weight, and a skinfold estimation of body fat composition. From these, a variable representing the lean body weight for each officer was calculated (total body weight − [percentage of body fat × total body weight]). This last measure is generally regarded as a proxy for size. These YMCA measures were available for the three previous annual assessments. The measures from the most recent assessment for each person were used. Test–retest correlations were calculated for those individuals who had multiple-year measurements on these variables. The relevant correlations were .87 for maximum oxygen uptake capacity and .70 for body fat composition, which are quite respectable given that some of the measures were separated by 2 years.

File rating. A rating was taken from one item on the police department's standard annual performance evaluation form. This item was "Control of Conflict—Voice Command/Physical Skill," evaluated on a 7-point scale (7 = *superior* and 1 = *unacceptable*). This rating was made by the officer's direct supervisor and is part of the yearly formal evaluations made for personnel purposes. This item served as a check against the physical ratings designed specifically for this study. It was significantly correlated with the (independently made) physical ratings of wrestling (.45), lifting and carrying (.48), pushing and pulling (.44), and overall job performance (.31).

Step 4: Sample

From the population of 448 police officers who were not on any form of limited status due to injury or illness, 200 were randomly selected to participate in this study. Of the random sample of 200, we were able to collect physical test data from 132 officers, and ratings from 167. This resulted in a final sample size of 115 officers (96 men and 19 women) for whom we had both test scores and ratings with which to test the fit of our model. The mean age of this sample was 35.4 years ($SD = 8.8$). There were no significant differences between the 115 officers included in the sample and the 85 officers excluded from further analyses because of missing data on the following variables: gender, age, maximum oxygen uptake, body fat composition, and lean body weight. The subjects in our final sample of 115 were representative of the larger sample of 200, which, of course, was a randomly drawn sample from the population of 448. Of these 115 officers, 89 had complete physiological data from the YMCA testing and 102 had file-rating information.

Table 3 presents the means, standard deviations, and pairwise correlations between each of the variables in the final sample of 115 officers. The correlations shown were used as the input matrices for the LISREL VII analyses discussed later. Estimates of the reliabilities for the test events are also presented in Table 3. The estimates shown are the communalities of the variables drawn from a principle factor analysis of the variables and, according to Harmon (1967, p. 19), represent conservative estimates of the reliabilities.

Step 5: Specification and Testing of a Confirmatory Factor Analysis Model

The set of correlations between the eight physical ability tests and the eight physical task performance ratings might be construed to constitute evidence for the validity of the tests, in the context of a concurrent validation strategy. However, because of potential problems with these ratings in terms of representing criterion variables without bias (see later discussion), we pursued a construct validity approach with the objective of confirming a model in which specific constructs were hypothesized as the organizing factors explaining much of the variance in test scores and ratings. Specifically, we proposed that two latent variables—strength and endurance—operate as the prime determinants for both test scores and ratings and produce the set of correlations among them. A diagram of our hypothesized model is shown in Figure 1.

Our model included a third latent variable—a ratings factor—which was expected to produce common variance among all of the ratings variables. This common variance factor can be thought of as method variance or as halo that extends to all ratings and represents that portion of the common variance among ratings that is not correlated with any other factor. We specified the paths from the ratings factor to each ratings variable to be equal (using the logic that method variance or halo is not, by definition, differentiated across rating dimensions), and we specified that the correlations of the ratings factor with strength and endurance be zero. Here we followed the precedent taken by Campbell, McHenry, and Wise (1990), who defined the method factor as that portion of the common variance among measures from the same method that is not predictable from any of the other related factors (i.e., strength and endurance).

To test our hypothesized model, we used LISREL VII (Jöreskog & Sörbom, 1989) to evaluate the model's fit to the data generated from sample information. The input to the program consisted of a matrix of correlations among the first 16 measured variables shown in Table 3. The fit of a hypothesized model indicates the likelihood that it could have produced the observed data. Goodness of fit can be assessed with several statistics, including an overall chi-square, the root-mean-square residual (rmsr), and the goodness-of-fit index (GFI). Also calculated was the ρ statistic developed by Bentler and Bonett (1980), who asserted that ρ values greater than .90 are generally acceptable.

LISREL VII also provides modification indices, which indicate the expected overall improvement in the chi-square fit measure resulting from freeing fixed parameters. Model modifications constitute a specification search to improve the fit of the model to the data. Minor modifications were made to the original model, though only where such changes could be theoretically justified. The modifications made to the initial model involved allowing LISREL to estimate correlations between the unique variances, or errors, for two pairs of measured variables instead of fixing them at zero. The correlation between the error components for ratings of climbing and of crawling and balancing was estimated at .103; the error correlation between dummy wrestle and dummy drag test scores was estimated at .153. Each of these was statistically significant at an alpha level less than .05. The implication is that there may be some influence on each of these variable pairs that is not accounted for in the model (e.g., some unmeasured variable). Based on these modifications and with these parameters freed, the resulting figures from the LISREL analyses were as follows: $\chi^2(96) = 126.87$, $p < .05$, rmsr = .060, GFI = .88, and $\rho = .97$. Although the chi-square was significant, the other indices indicated a quite acceptable fit.

Table 4 contains the standardized factor loadings for the model (as modified). The loadings generally confirm prior expectations about the relationships among the measured variables and the constructs of strength and endurance. Inspection of these values indicates that the variable most representative of the strength factor was grip strength, followed by lifting and carrying (a performance rating variable). Also, the dummy wrestle and dummy drag tests showed high loadings on this factor. Similarly, the obstacle course, 1-mile run, 100-yard dash, sit-up, and bench dip variables exhibited strong loadings on endur-

Table 3
Means, Standard Deviations, and Intercorrelations for Measured Variables in the Incumbent Officers Sample

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1. Grip strength	59.3	11.7	(.65)	114	114	110	113	114	113	107	114	115	115	115	115	115	115	115	89	101	115	88	88	115	115
2. Dummy wrestling	23.4	7.3	389*	(.61)	113	109	112	113	113	106	113	114	114	114	114	114	114	114	88	100	114	87	87	114	114
3. Dummy drag	11.3	3.1	354*	554*	(.73)	110	113	113	112	107	113	114	114	114	114	114	114	114	88	100	114	87	87	114	114
4. 100-yard dash	14.1	2.1	099	483*	486*	(.71)	110	110	108	105	109	110	110	110	110	110	110	110	84	96	110	83	83	110	110
5. Obstacle course	34.4	11.3	159*	508*	447*	545*	(.67)	112	111	106	112	113	113	113	113	113	113	113	87	99	113	86	86	113	113
6. Sit-ups	34.0	10.7	−118	249*	207*	410*	497*	(.69)	112	106	113	114	114	114	114	114	114	114	88	100	114	87	87	114	114
7. Bench dips	30.8	13.3	073	260*	176	387*	499*	470*	(.60)	105	112	113	113	113	113	113	113	113	87	99	113	86	86	113	113
8. 1-mile run	663.6	149.8	−067	247*	302*	462*	535*	562*	502*	(.64)	106	107	107	107	107	107	107	107	81	93	107	80	80	107	107
9. Wrestling	3.63	0.90	505*	305*	169	237	195*	091	140	109	—	114	114	114	114	114	114	114	88	101	114	87	87	114	114
10. Lifting and carrying	3.65	0.89	501*	328*	286*	208*	184	036	027	098	852*	—	115	115	115	115	115	115	89	101	115	88	88	115	
11. Pushing and pulling	3.70	0.93	467*	236*	232*	225*	201*	085	118	144	872*	880*	—	115	115	115	115	115	89	101	115	88	88	115	115
12. General physical fitness	3.63	0.89	094	198*	199*	369*	434*	375*	389*	384*	515*	509*	538*	—	115	115	115	115	89	101	115	88	88	115	115
13. Endurance	3.54	0.81	239*	166	240*	283*	286*	218*	252*	338*	560*	532*	594*	707*	—	115	115	115	89	101	115	88	88	115	115
14. Running	3.37	0.88	061	263*	258*	380*	502*	425*	313*	427*	455*	398*	426*	765*	669*	—	115	115	89	101	115	88	88	115	115
15. Climbing	3.47	0.82	098	235*	238*	401*	424*	371*	319*	430*	533*	515*	512*	721*	700*	731*	—	115	89	101	115	88	88	115	115
16. Crawling and balancing	3.46	0.82	033	192*	167	398*	398*	402*	324*	351*	467*	427*	474*	717*	614*	700*	798*	—	89	101	115	88	88	115	115
17. Maximum oxygen uptake capacity	41.4	7.1	062	368*	390*	549*	495*	584*	534*	681*	292*	247*	295*	564*	445*	551*	584*	566*	—	87	89	88	88	89	89
18. File rating	5.49	0.89	374*	177	207*	035	−019	−128	024	−045	451*	480*	442*	179	119	020	119	030	070	—	101	86	86	101	101
19. Overall job performance	3.56	0.77	336*	305*	172	251*	186*	144	117	181	670*	640*	713*	519*	595*	406*	524*	518*	348*	310*	—	88	88	115	115
20. Lean body weight	150.0	27.7	740*	259*	102	−215*	−149*	−382*	−170	−328*	434*	432*	327*	−111	014	−089	−114	−212*	−184	310*	098	—	88	88	88
21. Body fat composition	21.5	5.4	−123	−408*	−335*	−381*	−448*	−227*	−399*	−310*	−362*	−335*	−322*	−498*	−330*	−510*	−562*	−561*	−526*	−138	−325*	−003	—	88	88
22. Male	0.83	0.37	699*	463*	355*	237*	217*	−122	126	016	571*	513*	487*	103	240*	158	170	108	157	336*	382*	690*	−391*	—	115
23. Age	35.4	8.8	228*	−144	−215*	−398*	−449*	−565*	−240*	−295*	124	161	058	−174	−068	−230*	−177	−219*	−395*	361*	043	422*	−060	267*	—

Note. Correlations are presented below the diagonal (decimals omitted); pairwise *N*s are presented above the diagonal. Means and standard deviations are in raw units (grip strength in pounds; dummy wrestle, dummy drag, 100-yard dash, obstacle course, and 1-mile run in seconds; and sit-ups and bench dips in number per minute). Signs for correlations are for timed tests. Figures in parentheses represent the reliabilities of the variables, estimated from the commonalities obtained in a principle-components factor analysis (Harmon, 1967).

* $p < .05$.

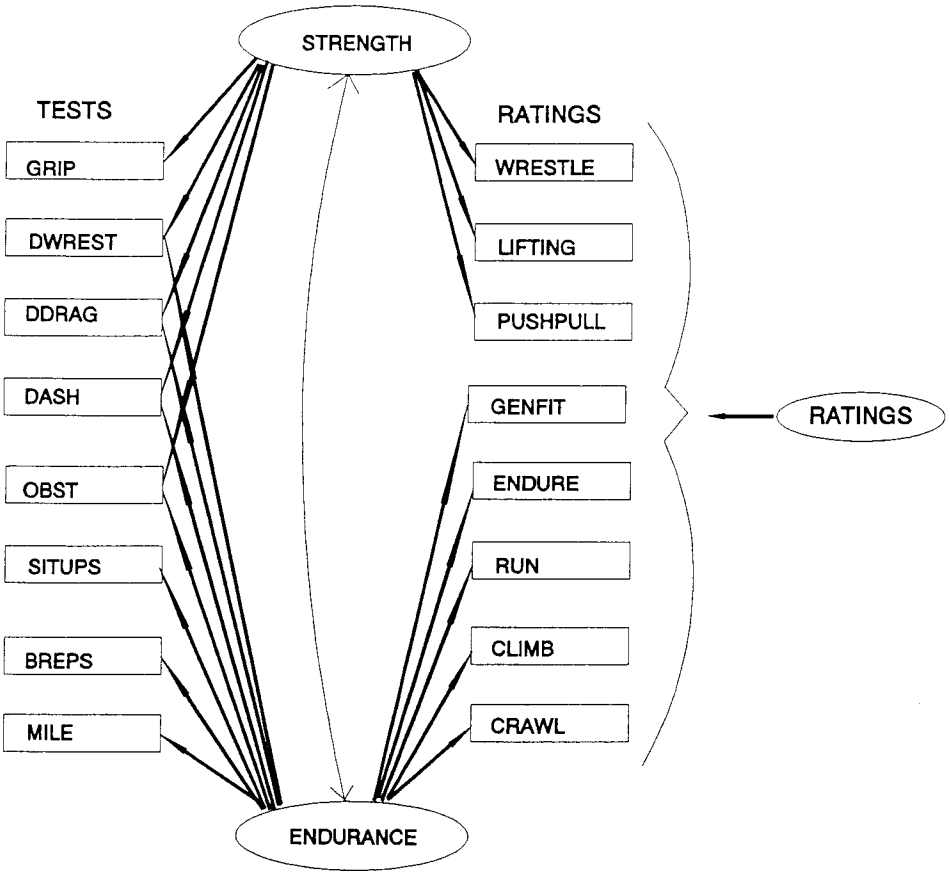

Figure 1. Hypothesized model of two latent variables—strength and endurance—with test and performance ratings (all performance ratings load equally on the latent ratings factor; GRIP = grip strength; DWREST = dummy wrestling; DDRAG = dummy drag; DASH = 100-yard dash; OBST = obstacle course; SITUPS = sit-ups; BREPS = bench dips; MILE = 1-mile run; WRESTLE = wrestling; LIFTING = lifting and carrying; PUSHPULL = pushing and pulling; GENFIT = general physical fitness; ENDURE = endurance; RUN = running; CLIMB = climbing; CRAWL = crawling and balancing).

ance. Dummy drag and dummy wrestle were about equally loaded on both strength and endurance. In general, the physical ability test events showed slightly higher loadings on endurance than on strength.

These data can also be reviewed in accordance with our stated expectations or hypotheses concerning whether each test would show a moderate or high loading on the two constructs (where no loadings were expected, we fixed these parameters at zero). We considered values of .60 and above to mean a relatively high loading and values of .30 to .60 to signify a moderate loading. Table 5 shows the correspondence between the parameter estimates and our hypothesized loadings. As can be seen, the estimated parameter values for the endurance factor are in accordance with our expectations, with one exception (the dummy drag event exhibited a moderate rather than a high loading). The estimated loadings for the strength factor were not as clear. As expected, the grip strength event was hypothesized to load highly on the strength factor and, in fact, it did (.659). However, the other four test events showed only moderate and low loadings on this factor when we expected relatively higher loadings. As a whole, however, there appears to be fairly good support for the a priori predictions made.

Much of the covariance among the ratings items is due to a common ratings factor; each rating had a .664 loading on this factor. However, after this common method variance was accounted for, it appears that supervisors could indeed discriminate between officers' strength and endurance factors, as we expected. Supervisors' ratings of wrestling, lifting and carrying, and pushing and pulling all loaded highly on strength, whereas their ratings of general physical fitness, running, climbing, and crawling and balancing loaded moderately on endurance. An unexpected result was the endurance rating, which proved to have only a moderate .331 loading on the endurance factor. A close reading of the verbal anchors used for this item suggests that raters may have interpreted it to reflect an officer's tolerance for difficult working conditions (i.e., heat, cold, and rain) rather than capacity for extended physical exertion.

Overall, the results of this analysis support our theoretical model about the role of strength and endurance as the prime determining factors affecting both physical test scores and performance rating measures. The overall fit of the model is adequate, and the hypothesized relationships between the test events and the constructs are generally supportive, although there are some exceptions.

Step 6: Replicating Part of the Model With an Applicant Sample

Because this research was conducted with the objective of using these tests operationally, the tests were administered to applicants. This administration provided us with an opportunity to evaluate the

Table 4
Factor Loadings for Standardized Solution of the Confirmatory Factor Analysis Model

Manifest variable	Latent variable: Strength	Endurance	Ratings	Unique variance
Grip strength	.659			.566
Dummy wrestle	.488	.442		.540
Dummy drag	.386	.429		.647
100-yard dash	.236	.645		.510
Obstacle course	.216	.778		.328
Sit-ups		.611		.627
Bench dips		.659		.566
1-mile run		.688		.527
Wrestling	.608		.664	.153
Lifting and carrying	.651		.664	.121
Pushing and pulling	.610		.664	.126
Endurance		.342	.664	.359
General physical fitness		.501	.664	.232
Running		.574	.664	.264
Climbing		.505	.664	.272
Crawling and balancing		.480	.664	.339

Note. Blank spaces indicate parameters fixed at zero. The correlation between strength and endurance was estimated at .061. Correlations between error terms were estimated for climbing and crawling and balancing (.103) and between wrestling and dummy drag (.153). All estimated parameters are significant at $p < .05$, except the correlation between strength and endurance.

model with data from an independent sample of individuals. The eight physical tests were administered to 161 actual applicants for the police officer position. The tests were administered in the same location, order, and manner as for the incumbent sample. The means, standard deviations, and test intercorrelations for this sample are given in Table 6. Also indicated in Table 6 are those test events on which the applicant sample differed significantly from the incumbent sample. These were the 100-yard dash, obstacle course, sit-up, and 1-mile run variables, on which the applicant sample scored relatively better than the incumbent sample, probably due to age, which also differed significantly across the two samples. The incumbent sample scored significantly better on the dummy drag test.

However, because there were no supervisory ratings for the applicant sample, only that portion of the model relating strength and endurance to the physical test scores could be replicated. The model for comparison, then, involved eight manifest variables and two latent variables, involving the paths found in the lefthand portion of Figure 1. The model was tested on the applicant sample, in which the values of particular loadings were fixed (based on the incumbent sample) but the error variances were estimated. This procedure corresponds to what Bentler (1980) called a moderate replication, an approach he favored in his *Annual Review of Psychology* chapter. The tests of model fit obtained for the applicant sample data are as follows: $\chi^2(28) = 52.08$, $p < .01$; GFI = .924, rmsr = .087, and $\rho = .952$. These values again indicate a very good fit of the data to the specified model. Thus, the parameter values and the overall model based on the incumbent sample appear to replicate in an actual applicant sample.

Step 7: Developing and Testing the Nomological Network With Exogenous Variables

The next step in our analysis was to extend the model by including a variety of other anthropometric and performance measures for incumbent officers. This follows the logic of construct validation—to explore the nomological network of constructs associated with strength and endurance. To this end, we formed hypotheses about how strength and endurance were related to seven other variables (maximum oxygen uptake, lean body weight, body fat composition, gender, age, and the two global rating items) and then examined whether the data supported these hypotheses.

Our hypotheses were based on research from the field of exercise physiology, as reported by McArdle et al. (1981), and were as follows:

1. Tests of maximum oxygen uptake are often used as direct measures of aerobic capacity, so we expected this variable to be highly correlated with endurance. However, nothing in the literature led us to expect any more than a low positive correlation with strength.
2. Lean body weight is a measure of body weight, or size, independent of body fat. It has a strong linear relationship to the volume and density of muscle possessed by an individual. Because muscular strength has a high linear relationship with the cross-sectional area of the muscular tissue used to perform a task (McArdle et al., 1981), we hypothesized a high positive correlation between lean body weight and strength. We did not expect any substantial association between lean body weight and endurance.
3. We predicted a high negative correlation between endurance and body fat composition, but only a low negative correlation between body fat composition and strength.
4. Prior research indicates that there is little difference in muscle strength when men and women have the same muscular mass size. However, research has established that women possess an average of 56% of the upper-body isometric strength and 72% of the isometric leg strength of men (McArdle et al., 1981). Research also indicates that there is a gender difference in aerobic capacity as well, although the gap is much smaller than for strength. The aerobic capacity difference is partly due to women having a higher percentage of body fat and a lower concentration of hemoglobin than men (McArdle et al., 1981). Because of these prior research findings, we hypothesized that gender would be strongly related to strength but would have only a low relationship with endurance.
5. Maximal aerobic power declines with age after the mid-20s to about 70% of peak at age 65 (McArdle et al., 1981). Muscular strength declines with age as well, but the rate of decline is slower and varies according to the specific muscles involved. On the basis of these past research findings, we expected the data to show a moderate negative correlation between age and endurance and a low correlation between age and strength.
6. We had little a priori basis for developing specific hypotheses concerning the relationships of the two global performance ratings with strength and endurance other than the rationale extended earlier

Table 5
Hypothesized Test Event Loadings and Estimated Parameters Provided by LISREL VII

Test event	Predicted loadings: Strength	Endurance	Estimated loadings: Strength	Endurance
Grip strength	H	—	.659	—
Dummy wrestle	H	M	.488	.442
Dummy drag	H	M	.386	.429
100-yard dash	M	M	.236	.645
Obstacle course	M	H	.216	.778
Sit-ups	—	H	—	.611
Bench dips	—	H	—	.659
1-mile run	—	H	—	.688

Note. H = hypothesized loadings .60 and above; M = hypothesized loadings from .30 to .60.

Table 6
Means, Standard Deviations, and Correlations Between Physical Tests in the Applicant Sample

Variable	*M*	*SD*	1	2	3	4	5	6	7	8
1. Grip strength	59.2	10.0	—							
2. Dummy wrestle	22.8	4.9	.290*	—						
3. Dummy drag	13.2[a]	2.5	.428*	.456*	—					
4. 100-yard dash	13.2[a]	1.3	.346*	.414*	.512*	—				
5. Obstacle course	27.6[a]	3.8	.255*	.499*	.539*	.833*	—			
6. Sit-ups	36.6[a]	8.2	.037	.196*	.194*	.391*	.396*	—		
7. Bench dips	32.6	10.7	.253*	.298*	.243*	.483*	.478*	.490*	—	
8. 1-mile run	552.5[a]	98.9	.107	.386*	.346*	.442*	.537*	.391*	.414*	—

Note. *N* for all variables is 161. Signs for correlations are reversed for timed tests.
[a] There is a significant difference between the means of the applicant and incumbent samples ($p < .05$, based on a pooled variance estimate).
* $p < .05$.

that strength and endurance should be related to independently gathered measures of overall job performance. Thus, we hypothesized that the two ratings would each be moderately related to both the strength and endurance factors. A summary of these predictions is given in the lefthand portion of Table 7.

To empirically test our hypotheses, we used LISREL VII to estimate the correlation of the seven variables with strength and endurance. The model specified the strength, endurance, and ratings factors to be related to the eight physical tests and eight ratings, exactly as before. Factor loadings, unique variances, correlations between the two error terms, and the correlations among the strength, endurance, and ratings factors were all fixed at values equal to the estimates shown in Table 4. Seven new latent variables were added to the model, each determining a single manifest variable (the unique variances were fixed at zero). The input matrix consisted of the entire correlation matrix for all 27 variables shown in Table 3. The only values estimated were the correlations among the seven new latent variables, and the correlations of these seven variables with the strength, endurance, and ratings latent variables. These correlations are shown in Table 8.

The righthand portion of Table 7 provides the level of correlations for these seven variables with strength and endurance, and one can simply compare these values against the predictions made by reviewing the right- and lefthand portions of this table.

Table 7
Hypothesized Correlations Between Strength and Endurance Factors and Other Variables, and Actual Estimates Provided by LISREL

	Predicted correlations		Estimated correlations	
Variable	Strength	Endurance	Strength	Endurance
Maximum oxygen uptake capacity	—	H	.120	.755
Lean body weight	H	—	.623	−.366
Body fat composition	—	H	−.186	−.517
Gender	H	—	.737	.100
Age	—	M	.079	−.562
Overall job performance	M	M	.389	.153
File rating	M	M	.501	−.138

Note. H = hypothesized correlations .60 and above; M = hypothesized correlations from .30 to .60.

One surprise was found: the −.366 correlation between lean body weight and endurance. One explanation for this could be that there is a diseconomy of scale with respect to size in terms of the efficiency of moving one's own body mass. It does not seem unreasonable, for example, to expect a 110-lb gymnast to do more pull-ups than a 250-lb football tackle, even if both have the same proportion of fat and muscle.

The last two variables in Table 7 are the two global rating items. The results show that each of these was moderately correlated with strength but lowly (and even negatively) correlated with endurance. Apparently, supervisors evaluated the strength components of the job as being of greater importance than the endurance components of the job.

In summary, the correlations among the latent variables in the expected model provide evidence that the constructs we identified as strength and endurance generally correlate with other variables, in correspondence with our expectations based on prior findings in exercise physiology.

Step 8: Examining Gender Differences on the Physical Tests

Because of the legal sensitivities identified in the Introduction, the next step in this process was to examine the data for potential bias against women. To this end, we first examined the data for significant differences between male and female incumbents. Table 9 presents the relevant means, standard deviations, and *t* tests used to determine if such differences exist on the various measures we had collected, as well as the factor scores for strength and endurance estimated from the LISREL analyses. Also shown in Table 9 are the relevant *z* score differences (the differences between the group means divided by the average of the standard deviations for the two samples) between the men and women.

These data indicate substantial differences between men and women on these measures. Men's scores differed significantly from women's on five of the eight physical ability test events. On these particular five tests, men scored better in the sense of performing the tests faster or showing greater physical strength. Moreover, there were also corresponding differences between men and women on the job performance ratings generated by supervisors. On four of the eight scales, men were rated significantly higher than women. However, no differences were observed on the ratings of general physical fitness, running, climbing, or crawling and balancing.

In terms of the general constructs of strength and endurance, men demonstrated significantly higher factor scores than women on the strength but not on the endurance factor.

Table 8
Correlations Between Strength, Endurance, and Ratings Factors and Seven Additional Variables

Variable	1	2	3	4	5	6	7	8	9	10
1. Strength	—									
2. Endurance	.061	—								
3. Ratings	.000	.000	—							
4. Maximum oxygen uptake capacity	.120	.755*	.234*	—						
5. File rating	.501*	−.138	.208*	.054	—					
6. Overall job performance	.389*	.153*	.621*	.326*	.304*	—				
7. Lean body weight	.623*	−.366*	.050	−.198*	.311*	.093	—			
8. Body fat composition	−.186*	−.517*	−.297*	−.524*	−.125	−.307*	.010	—		
9. Gender	.737*	.100	.092	.146	.332*	.369*	.687*	−.382*	—	
10. Age	.079	−.562*	.148*	−.401*	.365*	.057	.424*	−.056	.265*	—

* $p < .05$.

Thus, as expected, we observed major differences between men and women on strength-oriented measures but not on endurance-oriented measures. An important element to note here is that these differences were observed across the test events as well as across the criterion components of the job.

Table 9 also indicates that men and women differed on such variables as age, lean body weight, and body fat composition, which were important correlates with the constructs of strength and endurance (see Table 8). A question therefore arises concerning whether the test differentials observed between men and women could be a result of other variables. That is, perhaps there is nothing inherent in being male or female in producing these observed score differentials; instead these differences might be produced by or associated with other correlated variables. We therefore decided to explore whether any gender differentials would be associated with particular test events after a number of correlates were held constant. Similarly, we wanted to determine if any age differences would be associated with these test events after such variables were held constant.

Table 9
Differences Between Male and Female Incumbents

Variable	Men	Women	t^a	z^b
Age	36.4	30.2	2.94*	.074
Lean body weight	158.7	108.1	8.85*	2.51
Body fat composition	20.6	26.2	3.94*	1.12
Grip strength	62.9	40.9	10.38*	2.51
Dummy wrestle	22.5	31.5	5.52*	1.43
Dummy drag	10.8	13.7	4.02*	1.01
100-yard dash	13.9	15.3	2.54*	0.65
Obstacle course	33.3	40.0	2.34*	0.60
Sit-ups	33.4	36.9	1.30	−0.33
Bench dips	31.6	27.1	1.34	0.34
1-mile run	662.5	668.7	0.16	0.04
Wrestling	3.85	2.44	7.35*	1.89
Lifting and carrying	3.85	2.63	6.35*	1.59
Pushing and pulling	3.90	2.68	5.92*	1.50
General physical fitness	3.67	3.42	1.10	0.28
Endurance	3.63	3.10	2.62*	0.67
Running	3.42	3.05	1.71	0.42
Climbing	3.53	3.16	1.03	0.46
Crawling and balancing	3.50	3.26	1.15	0.29
Strength factor	0.27	−1.48	9.72*	2.58
Endurance factor	0.15	−0.24	1.66	0.44
Ratings factor	0.09	−0.44	2.11*	0.56

[a] Absolute values from a test of equality of means using a pooled variance estimate.
[b] Positive when in the men's favor, negative when in the women's favor.
* $p < .05$.

The question we decided to pursue was the following: "Are there test differences between men and women after pertinent factors expected to be associated with such tests are held constant?" A similar question can be raised about whether age differences remain after holding such variables constant.

To this end, we used hierarchical regression procedures in which variables were entered sequentially to predict test scores for each of the eight physical ability test events. The procedure first forced the strength and endurance factor scores into the regression equation in which scores on these two factors were estimated with LISREL procedures that excluded the particular physical test event score from the estimate. (Recall that the strength and endurance factors were derived from a composite of both physical test events and performance ratings. In each particular analysis, the specific test event was not included in the factor score for strength or endurance because to do so would mean that the same variable was included in both the independent and dependent variable space.) We elected to include these two factors as independent variables because we wanted to ascertain if there were test score differentials for different gender and age groups when strength and endurance were held constant. Subsequently, a forward stepwise procedure was used to allow entry of the physiological measures of lean body weight and body fat composition into the model to determine if these variables predicted the test events. This constituted the first regression model. A second regression model was developed which forced in the independent variables from the first two blocks and then used a forward stepwise criteria for the entry of gender and age. Table 10 presents the results from these analyses. The change in the squared multiple correlations can be examined to determine if age and gender were significantly related to test scores when the strength and endurance factors as well as lean body weight and body fat composition were held constant. Similarly, the significance of the beta weights can be examined to determine if age or gender were independently related to test score when these other factors were held constant.

The change in the squared multiple correlations indicates that gender and age were seldom significantly related to test scores after these other variables were partialed out. Gender was significantly related to the grip strength test after strength and lean body weight were

Table 10
Hierarchial Regression of Physical Factors, Lean Body Weight, Body Fat Composition, Gender, and Age on Physical Test Scores

Dependent variable	Strength factor		Endurance factor		Lean body weight		Body fat composition		Male		Age		
	β	t	β	t	β	t	β	t	β	t	β	t	R^2
Grip strength													
First model	.25	2.8	—	—	.60	6.7	—	—	—	—	—	—	.61
Second model	.15	1.6	—	—	.46	4.4	—	—	.29	2.6	N		.64
Dummy wrestle													
First model	.25	2.1	.53	5.7	.33	2.6	N		—	—	—	—	.50
Second model	.25	2.0	.46	4.3	.30	2.3	−.12	−1.3	N		N		.51
Dummy drag													
First model	.44	4.4	.26	2.7	N		N		—	—	—	—	.30
Second model	.36	2.4	.27	2.0	.09	0.6	−.06	−0.05	N		N		.31
100-yard dash													
First model	.33	2.5	.53	5.5	−.30	−2.2	N		—	—	—	—	.45
Second model	.32	2.4	.43	3.9	−.33	−2.5	−.19	−1.8	N		N		.48
Obstacle course													
First model	.22	2.8	.72	9.1	N		N		—	—	—	—	.55
Second model	.17	1.4	.52	5.0	.05	0.4	−.11	−1.1	N		−.22	−2.2	.58
Sit-ups													
First model	—	—	.59	5.7	—	—	N		—	—	—	—	.35
Second model	—	—	.36	2.8	—	—	.12	1.1	N		−.42	−3.9	.47
Bench dips													
First model	—	—	.56	5.7	—	—	N		—	—	—	—	.31
Second model	—	—	.53	4.6	—	—	−.06	−0.5	N		N		.32
1-mile run													
First model	—	—	.62	6.7	—	—	N		—	—	—	—	.38
Second model	—	—	.66	6.0	—	—	.08	0.7	N		N		.38

Note. Dashes indicate independent variables that were not included in the model, and *N*s indicate variables that were in the model but did not meet the criteria for entry in a forward stepwise procedure. The strength and endurance factor scores, as appropriate to the dependent variable, were forced into each model. A forward stepwise procedure was used in the first regression model for each dependent variable to allow entry of lean body weight, body fat composition, or both if they were significant. In the second regression model, the independent variables from the first two blocks were forced in and then forward stepwise criteria were used for entry of significant coefficients for gender (male) and age.

accounted for; men scored significantly higher or better than women. Gender was not significantly related to any of the other test events after the other variables were held constant. Age showed a significant correlation with scores on the obstacle-course and sit-up test events, on which older subjects performed less well, even when these other constructs and physiological variables were controlled for.

To further explore the issue of potential gender bias in a more traditional fashion, we decided to conduct regression analyses in which the test events, gender (with men coded as 1 and women coded as 0), and the interaction between gender and test were variables in the equations estimated when predicting strength- and endurance-oriented performance ratings. The question asked here was, "Does gender significantly correlate with job performance after test scores are held constant?" Readers familiar with the test fairness issue will recognize this perspective as representing Cleary's (1968) model of test fairness, whereas the previous formulation represents Cole's (1973) conditional probability model, which was more recently articulated as the model adopted by the National Research Council (Hartigan & Wigdor, 1989) for examining bias in the General Aptitude Test Battery.

To this end, we developed two job performance composites (reflecting the strength and endurance factors) from supervisors ratings. Composite strength and endurance performance measures were formed by simply summing across those supervisory performance ratings shown to load highly on the strength and endurance factors after standardization. Thus, a composite strength-oriented performance factor was formed by summing across the performance ratings for wrestling, lifting and carrying, and pushing and pulling. Similarly, a composite endurance-oriented performance rating was formed by summing across the supervisors' ratings of running, endurance, climbing, general physical fitness, and crawling and balancing.

Regression equations were developed in which the strength- and endurance-oriented performance ratings served as separate dependent variables and three independent variables for each test event (the specific test event, gender, and the interaction of test and gender) were entered into the equation. Significant beta weights for the interaction term signify that differential slopes exist for men and women. When there was a nonsignificant interaction or no slope differential, a significant beta weight for the gender term indicated differential intercepts; if a significant interaction or differential slopes were found, the interpretation of a significant gender effect became more complex because of the differential slopes of the regression lines for men and women. Finally, a significant beta for the test event indicated that, when both intercept and slope differences were held constant, the test event was still significantly correlated with that particular aspect of performance. These results are shown in Table 11, along with the particular slope values obtained for men and women when the regression line was estimated for each gender separately.

Reviewing the data concerning the endurance performance rating first, results indicate that with one exception there were no significant slope differentials for men and women on these test events. For the bench-dip test event, however, results indicated that the data for men and women exhibited different regression slopes in the prediction of this dependent variable. As can be seen, the slope value was .25 for men and .83 for women. This particular test exhibited a higher relationship

Table 11
Regressions of Ratings on Tests

Independent variable	Full sample		Male sample		Female sample		Intercept difference[a]	
	β	*SE*	β	*SE*	β	*SE*	β	*SE*
Regression of standardized strength ratings composite on standardized strength-related physical test scores								
Grip strength[b]	.52	.08	.18	.10	1.08	.45	−0.23	.58
Dummy wrestle	.30	.09	.07	.09	−.03	.30	1.52	.32
Dummy drag	.24	.09	.05	.07	.05	.33	1.46	.28
100-yard dash	.24	.09	.07	.08	.25	.28	1.39	.25
Obstacle course	.20	.09	.05	.08	.26	.28	1.40	.24
Regression of standardized endurance ratings composite on standardized endurance-related physical test scores								
Dummy wrestle	.24	.09	.15	.11	.32	.29	.17	.33
Dummy drag	.26	.09	.24	.10	.10	.36	.39	.32
100-yard dash	.43	.09	.43	.09	.27	.29	.37	.26
Obstacle course	.46	.08	.46	.09	.36	.29	.30	.25
Sit-ups	.40	.08	.44	.08	.39	.30	.67	.29
Bench dips[b]	.34	.09	.25	.09	.83	.33	.31	.24
1-mile run	.45	.09	.46	.09	.34	.33	.51	.23

Note. The dependent variable for the strength regressions was a standardized sum of the wrestling, lifting and carrying, and pushing and pulling ratings. The dependent variable for the endurance regressions was a standardized sum of the general physical fitness, endurance, running, climbing, and crawling and balancing ratings.

[a] Intercept difference between the male and female samples. When positive, this value represents overprediction for women at the mean of the test score.

[b] Regression slopes for the male and female samples differed significantly ($p < .05$).

with rated performance for women than for men, although the test was significantly correlated with performance for both groups. Only the sit-up test event exhibited significantly different intercept values for men and women, as signified by a significant beta weight for gender. Being male led to a higher level of predicted endurance than did being female, given the same test score. However, if a common regression line were used, women would be overpredicted in the sense that their predicted performance scores would be higher than if gender had been used in the equation, Thus, although some bias appears to be manifested when this particular test event is examined against this endurance-oriented performance rating, the use of a common regression line or a top-down scoring system would not disadvantage women. A similar intercept difference was exhibited on the 1-mile run test event but, as with the sit-up event, women are not likely to be disadvantaged. These findings generally parallel the research associated with cognitive ability testing, in which the performance of the group with the lower intercept is overpredicted (Arvey & Faley, 1989). The other test events predicted the endurance performance measure rather well, with no slope or intercept differences exhibited.

Reviewing the results for the strength-oriented performance ratings yielded more complex results. These data indicate that differential slopes for men and women were obtained for the grip strength test event. In this instance, the test was more highly related to the dependent variable for women ($\beta = 1.08$) than for men ($\beta = .18$). In all other test events, significant intercept values were observed. However, women would be overpredicted if a common regression line or top-down selection strategy were used.

Thus, these results again suggest that there may be some problem with the grip strength test. The performance of men and women differed dramatically on this test, yet the test shows only a marginal relationship with the strength-oriented performance ratings for men and a more significant relationship for women.[2] Similarly, earlier analyses indicated that there were gender differences on this test after physiological and performance-oriented factors were controlled. Thus, if this test is adopted for selection purposes in the future, separate regression lines for men and women would be the most appropriate strategy to use in predicting performance.

A further check on whether gender bias is manifest in these data could be to again utilize LISREL methods (Arvey, Maxwell, & Abrahams, 1985). However, the sample size of the female cohort used in the present study was simply too limited ($n = 19$) for us to do any analyses with sufficient power. We did, however, re-estimate the measurement and structural models presented earlier, using only the male sample, to determine if the construct validity portion of our study would be supported in such a within-gender analysis. Using a reduced sample size of 96 men, we re-estimated the LISREL measurement model provided in Figure 1. The resulting values were as follows: $\chi^2(96) = 122.41$, $p < .05$, GFI = .87, rmsr = .071, and $\rho = .98$. Thus, the general construct validity of the model appears to be substantiated for this male-only sample. Examination of the test loadings indicated, however, that the test loadings on the strength factor were substantially lower than for the full sample. These diminished values are probably due to the restriction in range that occurred on almost all variables when the female sample was excluded. To check on this possibility, we corrected the correlations for range restriction (correcting for doubly truncated

[2] However, the grip strength test showed a stronger relationship for men than any other strength-oriented test and showed the highest overall relationship for men and women combined.

correlations), using the full-sample standard deviations for the unrestricted estimates of the variances (as calculated with the approximation procedure given by Alexander, Carson, Alliger, & Carr, 1987). These estimates were resubmitted to the same LISREL analysis. In accordance with our hunch, the tests specified to load on the strength factor showed substantially higher loadings (e.g., the grip strength test showed a .491 loading after correction, compared with a .274 loading without the correction). However, the fit of the overall model based on these corrected correlations was not as good, $\chi^2(96) = 259.24$, $p < .01$, GFI = .78, rmsr = .094. Perhaps the approximations obtained with this procedure produced some anomalies in the correlation matrix (see Hunter & Schmidt, 1990, pp. 50–52, for their remarks regarding the accuracy of this procedure).

We elected not to conduct separate model tests for high and low age groups because the sample sizes would be reduced sufficiently to make interpretation difficult. However, such analyses would be interesting given larger sample sizes.

Discussion

This study provides fairly convincing evidence for the construct validity of a set of physical ability test events that can be used in selecting entry-level police officers. The evidence suggests that two constructs—strength and endurance—underlie both performance on the tests and performance on the job, thus satisfying the linkage requirements for construct validity articulated by the Uniform Guidelines on Employee Selection Procedures (1978) as well as supporting the nomological network of relationships we developed from theory to establish the construct validity of such tests. Test and performance data obtained from a sample of incumbent officers fit the construct model well, as did data gathered from job applicants. Finally, regression analyses showed that women would be overpredicted if a common regression line were used for most tests and that data obtained from only the men in the incumbent sample also fit the construct model. We also confirmed our a priori expectations about the pattern of relationships among strength and endurance and such variables as sex, age, maximum oxygen uptake capacity, body fat composition, and lean body weight.

One test that did not appear to be a particularly good test in terms of potential gender bias was the grip strength test. The problem with this test was that gender was significantly related to test scores after the strength factor score and lean body weight had been partialled out. Moreover, the grip strength regression coefficients were different for men and women when predicting strength-oriented performance ratings. If this particular test is adopted for selection purposes, separate regression lines should be established for men and women for use in the selection process. Another obvious approach would be to search for alternative predictors that have similar loadings on the strength factor but have less of an adverse impact. This is, of course, a feature of the Uniform Guidelines that makes good sense in this context, notwithstanding the difficulty of finding an equally fair predictor with less adverse impact (cf. Maxwell & Arvey, in press). Tests that assess lower-body muscular strength (Hogan, 1991a) are good candidates. An interesting issue surfaces here, however. The job analysis data (i.e., the Use of Force Reports) were quite convincing regarding the use of upper-body strength when dealing with arresting or subduing suspects, and so forth. Thus, substituting measures of lower-body strength for measures of upper-body strength may conflict with actual job requirements, at least according to the job analytic evidence collected here. However, we noted earlier that the job analysis information gathered for this study should not be considered determinative.

There are several limitations to this study. One of the issues concerns whether the supervisors were sufficiently sensitive to the physical performance levels of the officers and whether they were simply responding to stereotypes based on body size and body fat when making their performance ratings. A similar issue can be raised about whether supervisors biased their ratings on the basis of gender. These are difficult and intractable problems with data of these sort. However, the data show several relevant outcomes which suggest that the variance produced from these ratings are, indeed, valid: (a) The pattern of correlations among the ratings indicate some differentiation, even after method variance or halo was partialed out; (b) the ratings showed significant correlations with an independently collected and non–research-based performance rating drawn from personnel files; and (c) data based on these ratings fit the construct model quite well. If these ratings were substantially biased or were composed of sizable amounts of error variance, one would not expect the coherency found among the relationship patterns observed.

A second issue concerns whether the construct model developed was sufficiently complex. Might not a model that includes flexibility be more comprehensive? Two problems precluded our inclusion of this construct in the model. First, the capability of LISREL analyses and estimates to achieve an adequate fit to a specified model is a function of the number of measured and latent variables and the ratio of subjects to variables. Including a third latent variable would have been problematic in this regard. Second, there were no measured variables that seemed, at face value, to directly capture this particular construct. Instead, the a priori specified loadings between the observed measures and such a latent variable would have been somewhat low to moderate. In retrospect, however, it seems reasonable to develop and utilize one or two measures to capture flexibility more directly.

We believe that further replication and verification of this and similar models should be undertaken. It is our belief that construct validation approaches have more applicability than have previously been realized. One consideration, however, is the cost of such efforts. The multiple measures that need to be collected and the extensive and sophisticated data analyses are such that such model construction and verification is time consuming and costly. The yield in terms of understanding seems to be an important outcome of such efforts and cannot always be estimated in terms of costs and resource utilization.

Finally, these data and results still do not answer some rather thorny problems associated with the use of physical ability tests in employment settings. We do not know, for example, the relative importance of physical abilities and cognitive abilities in determining job performance, and thus we do not have any information concerning the appropriate weights for such tests. Moreover, the job analysis information and test data provide little guidance concerning the appropriate standards or cutoff points to adopt when such tests become operational. In short, a

number of operational and theoretical issues remain to be explored.

References

Alexander, R. A., Carson, K. P., Alliger, G. M., & Carr, L. (1987). Correcting doubly truncated correlations: An improved approximation for correcting the bivariate normal correlation when truncation has occurred on both variables. *Educational and Psychological Measurement, 47,* 309–315.

Americans With Disabilities Act of 1990. §933, Public Law 101–336 (1990).

Arvey, R. D., & Faley, R. H. (1989). *Fairness in selecting employees* (2nd ed). Reading, MA: Addison-Wesley.

Arvey, R. D., Maxwell, S. E., & Abrahams, L. (1985). Reliability artifacts in comparable worth procedures. *Journal of Applied Psychology, 70,* 695–705.

Arvey, R. D., Nutting, S. M., & Landon, T. E. (1992). Validation strategies for physical ability testing in police and fire settings. *Public Personnel Management, 21,* 301–312.

Bentler, P. M. (1980). Causal modeling. *Annual Review of Psychology, 31,* 421–440.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588–606.

Bernardin, H. J. (1988). Police officer. In S. Gael (Ed.), *The job analysis handbook for business, industry and government* (Vol 2, pp. 1242–1254). New York: Wiley.

Booth, J. W., & Hornick, K. (1982). Physical ability tests for police officer. *Police Chief, 28,* 38–42.

Campbell, J. P., McHenry, J. J., & Wise, L. L. (1990). Modeling job performance in a population of jobs. *Personnel Psychology, 43,* 313–333.

Campion, M. A. (1983). Personnel selection for physically demanding jobs: Review and recommendations. *Personnel Psychology, 36,* 527–550.

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

Cole, N. S. (1973). Bias in selection. *Journal of Educational Measurement, 10,* 115–124.

Fleishman, E. A. (1964). *The structure and measurement of physical fitness.* Englewood Cliffs, NJ: Prentice-Hall.

Fleishman, E. A. (1988). Some new frontiers in personnel selection research. *Personnel Psychology, 41,* 679–701.

Harmon, H. H. (1967). *Modern factor analysis* (2nd ed.). Chicago: University of Chicago Press.

Hartigan, J. A., & Wigdor, A. K. (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery.* Washington, DC: National Academy Press.

Hogan, J. (1991a). Physical abilities. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial-organizational psychology* (2nd ed., pp. 753–831). Palo Alto, CA: Consulting Psychologists Press.

Hogan, J. (1991b). The structure of physical performance in occupational tasks. *Journal of Applied Psychology, 76,* 495–507.

Hogan, J., & Quigley, A. M. (1986). Physical standards for employment and the courts. *American Psychologist, 41,* 1193–1217.

Hoover, L. T. (1989, June). *Trends in police physical agility selection testing.* Paper presented at the International Personnel Management Association Assessment Council, Orlando, Florida.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Jeanneret, P. R., Moore, J. R., Blakey, B. R., Koelzer, S. L., & Menkes, O. (May, 1991). *Development and validation of trooper physical ability and cogniitve ability tests* (Final report submitted to the Texas Department of Public Safety). (Available from Jeanneret & Associates, 3223 Smith Street, Suite 212, Houston, Texas 77006–6685.)

Jöreskog, K. G., & Sörbom, D. (1987). *LISREL VII.* Mooresville, IN: Scientific Software.

Lewis, R. E. (1989). *Physical ability tests as predictors of job-related criteria: A meta-analysis.* Unpublished manuscript.

Maher, P. T. (1984). Police physical ability tests: Can they ever be valid? *Public Personnel Management, 13,* 173–183.

Maxwell, S. E., & Arvey, R. D. (in press). The search for predictors with high validity and low adverse impact: Compatible or incompatible goals? *Journal of Applied Psychology.*

McArdle, W. D., Katch, F. I., & Katch, V. L. (1981). *Exercise physiology.* Philadelphia, PA: Lea & Febiger.

Schmitt N. (1988, October). Construct validation in personnel selection research: Some alternatives. In *Proceedings of the 1988 National Assessment Conference.* (Available from Personnel Decisions, Inc., 20000 Plaza VII Tower, 45 S. Seventh St., Minneapolis, Minnesota 55402.)

Uniform Guidelines on Employee Selection Procedures. (1978). *Federal Register, 43,* 38290–38315.

Vance, R. J., Coovert, R. D., MacCallum, R. C., & Hedge, J. W. (1989). Construct models of task performance. *Journal of Applied Psychology, 74,* 447–455.

Received October 15, 1991
Revision received March 2, 1992
Accepted March 19, 1992 ■