Construct Validity of Physical Fitness Tests

Ross R. Vickers, Jr.

Naval Health Research Center

Document No. 11-52

Naval Health Research Center
140 Sylvester Rd.
San Diego, California 92106-3521

Construct Validity of Physical Fitness Tests

Ross R. Vickers, Jr.

Naval Health Research Center

Warfighter Performance Department

San Diego, CA 92106-3521

e-mail: ross.vickers@med.navy.mil

Report No. 11-52 was supported by the Office of Naval Research, Arlington, VA, under Work Unit No. 60704. The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, or the U.S. Government. Approved for public release; distribution is unlimited. Human subjects participated in this study after giving their free and informed consent. This research has been conducted in compliance with all applicable federal regulations governing the protection of human subjects in research.

Abstract

The process of defining physical fitness test batteries typically relies on qualitative evaluations of individual tests. Starting from the existing consensus regarding the mapping of physical fitness tests onto physical ability constructs, analyses were carried out to develop quantitative test validity indices for use in test battery design. The validity indices were averaged factor loadings from confirmatory factor analysis (CFA) of the inter-test correlation matrices from 85 independent samples. The CFA included latent traits representing muscular strength, muscular power, muscular endurance, and cardiorespiratory endurance. The averaged factor loadings came from random effects analysis of the factor loadings from the 85 measurement models. The results confirmed the accepted assignment of fitness tests to categories representing the four physical ability constructs. The average factor loading varied from test to test within each category, but the inter-test variation generally was small relative to the standard errors of the individual loading estimates. The modest validity differences leave considerable freedom to use additional criteria, such as ease of administration, time requirements, and face validity from the perspective of the test population, when designing physical fitness test batteries.

Construct Validity of Physical Fitness Tests

Physical fitness tests measure physical abilities. Fitness test batteries frequently are used to assess the ability to meet occupational performance requirements. Valid tests must be selected for a battery to derive valid inferences about performance potential. Occupational requirements and time and equipment requirements influence test battery designs for specific applications. Expert judgment is the primary basis for choosing tests. The experts select tests that are believed to measure relevant physical abilities and that can be administered within the time and equipment constraints for testing. Expert judgment is largely qualitative. For example, judgments must be made regarding what can be measured. Judgment is needed because factor analyses have identified between 3 (Hogan, 1991) and 14 (Nicks & Fleishman, 1962) physical abilities or physical proficiency factors that can be measured by physical fitness tests. Judgments must be made about how many factors there really are and, of those, which ones are relevant to the current testing objectives. The factor analytic research also classifies fitness tests into groups representing different physical abilities. For example, push-ups and pull-ups are accepted muscular endurance (ME) measures, while a distance run is an accepted measure of cardiovascular endurance (CE). When the decision to measure a given ability is made, additional decisions are needed to decide which test or tests to use for making the desired measurements. The additional decisions are needed because the current state of the art provides little guidance for choosing among the tests that measure the particular abilities of interest.

This paper presents a reanalysis of a large volume of evidence relating physical fitness tests to physical abilities. The results provide a catalogue of quantitative construct validity coefficients for individual physical fitness tests. The catalogue entries rank individual physical fitness tests from most to least valid as indicators of the associated ability construct. The validity information can supplement expert judgment when designing fitness test batteries.

The catalogue covers four major physical ability constructs. Hogan's (1985) conceptual framework was the starting point for this effort. Hogan's framework consisted of seven physical ability constructs that

> …provide comprehensive coverage of the physical performance domain; the dimensions meet the following four criteria: (a) recognized research history; (b) definition consistent with human physiology; (c) measurement yielding variability across individuals; (d) association with performance in a variety of activities and tasks. (Hogan, 1985, p. 220)

This paper focuses on four of Hogan's (1985) constructs: muscular strength (MS), muscular power (MP), ME, and CE. These physical abilities were the focus because they appear most frequently in the job performance literature. Based on past practice, these constructs are likely to be of interest in designing occupational fitness test batteries.

Hogan's (1985) model derived from an extensive history of factor analyses of physical fitness tests. In this model, MS is "(t)he capacity to exert force as a result of tension produced in muscles." MP is "(t)he capacity to exert force to move a mass a given distance during a measured time." ME is "(t)he capacity of muscles to continue work over time while resisting fatigue." CE is "(t)he capacity of the heart and related body systems to sustain prolonged muscular activity." These four constructs are accepted in the physical fitness literature as factors that have been replicated across studies. Tests that are indicators of each construct have been identified in the replication process.

This study focuses on the problem of selecting specific tests for a physical fitness test battery. This study provides analyses that make it possible to choose tests based on their construct-related validity.

Method

Literature Search

The validity coefficients derived for physical fitness tests are based on the relationships among those tests. A literature search identified papers that provided correlation matrices describing those relationships. Previous reviews by Nicks and Fleishman (1962) and Fleishman (1964) were the search starting points. Computer searches coupled the four ability constructs with the keyword “measurement.” Variants of the construct names were employed. An ancestry search of papers that met the initial inclusion criteria identified additional relevant references.

The review was limited to studies that met minimum data requirements. The primary criterion was that the study had to report a correlation matrix that included at least three tests for a single physical ability. For example, a study was included if it reported all of the correlations between three or more ME tests. The criterion was relaxed slightly for studies that investigated two or more physical abilities. Those studies were included if the correlation matrix covered two or more tests for each of two or more abilities. These inclusion criteria ensured that the data would produced statistically acceptable measurement models. Sixty-eight studies that met the inclusion criteria reported results for 85 samples.

Demographics

The typical study participant was a man (Table 1). Roughly 50% of the samples and 50% of the total number of participants were men. Men contributed 47% of all test scores. The true contribution of men to the overall data probably is much larger. These figures excluded Blakly, Quinones, Crawford, and Jago’s (1994) study of 13,000 participants who provided 52,000 test scores.

The age–sex composition of the samples was noteworthy. The test data from male and female subjects were combined in the analyses for 14 of 17 samples of children and adolescents compared with 3 of 62 samples of adults.

Construct Sampling

The number of constructs represented in the measurement models varied from sample to sample (Table 2). Isometric MS, ME, MP, and CE measures were administered to between 36% and 52% of all samples. Isotonic MS and dynamic MS measures were administered to 8% and 13% of the samples, respectively.

The measurement models included as many as six latent traits even though the literature search focused on four physical ability constructs. The additional latent traits were added because different MS measurement methods produced distinct latent traits (Appendix A). Preliminary analyses demonstrated that although all strength tests measured the same general construct, the specific measurement method affected the representation of that construct. For example, the strength construct defined by a set of isometric strength tests was highly correlated with, but not identical to, the strength factor defined by a set of isoinertial strength tests. Neither of those factors was identical to the strength factor defined by a set of dynamic strength tests. This methodological variation meant that strength tests defined as many as three latent traits in some studies.

Adding three latent traits for MS to the latent traits for ME, MP, and CE meant that the ability measurement models could include as many as six latent traits. Few models were this complex. The measurement model represented just one construct in 36 samples. The model represented two constructs in 29 samples. The model represented three constructs in 13 samples. Four constructs were represented in the models for five samples. The model represented six constructs in only two samples.

Analysis Procedures

Model construction began by assigning tests to physical ability categories. Each test was assigned to a single category based on its usual interpretation in the testing literature. These assignments were straightforward except in the case of run tests. Shorter runs generally are classified as MP tests; longer runs are classified as CE tests. In the present case, run tests that covered 600 yd or less were classified as MP indicators; run tests that covered 880 yd or more were classified as ME indicators. This cutoff was based on Disch, Frankiewicz, and Jackson's, (1975) factor analysis of performance on run tests of 50 yd, 100 yd, 0.50 mi, .75 mi, 1.00 mi, 1.25 mi, 1.50 mi, 1.75 mi, 2.00 mi, and 12 min. The factor analysis produced two factors, one defined primarily by the shortest runs and one defined primarily by the longest runs. Intermediate runs of 0.50 mi to 1.0 mi had much larger loadings on the factor defined by long runs than on the factor defined by short runs.

The confirmatory factor analysis (CFA) model was unidimensional when all tests administered to a sample represented a single construct. The test battery had to include at least three tests, the minimum number required to identify a latent trait. Multidimensional CFA models were constructed when the correlation matrix included two or more tests for each of two or more ability constructs.

The measurement models combined free and constrained factor loadings for each physical ability test. A loading for each test on its hypothesized factor was freely estimated. The factor loadings for each test on all other factors were fixed at zero.

The remaining CFA model elements defined the latent trait structure for the models. All latent traits were scaled by fixing their variances at 1.00. This scaling choice made it possible to estimate factor loadings for all tests. Also, the latent trait correlations were freely estimated in the multidimensional CFA models. The analyses were conducted using LISREL 8 (Joreskog, Sorbom, du Toit, & du Toit, 2000).

Random effects (RE) meta-analyses estimated the average latent trait loadings (λ_{Avg}) for each test on its hypothesized ability dimension. The meta-analytic computations weighted individual latent trait loadings by the inverse of their variance. The variance was the squared standard error for the loading in the CFA model. This weighting scheme was adopted after preliminary analyses demonstrated that the CFA models produced appropriate standard errors even though correlation matrices were being analyzed (Appendix B).

An RE model was adopted to obtain results that could be generalized beyond the studies in the analysis (see Borenstein, Hedges, Higgins, & Rothstein, 2009). An RE model was appropriate because differences in participant characteristics (e.g., age, sex, general fitness), test setting (e.g., academic vs. military), and procedural differences in test administration (e.g., 1-min push-ups vs. 2-min push-ups) made it unlikely a priori that the factor loadings would be invariant across studies. Furthermore, RE analyses yield fixed effect models when there is little or no empirical variation in the parameter estimates. An SPSS-PC, version 17, syntax program to implement the procedures in Borenstein et al. (2009) was written and applied to conduct the analyses.

Results

Muscular Strength

A test was acceptable if λ_{Avg} was significantly, $p < .05$, greater than .40. This acceptability criterion was more stringent than the $\lambda = 0.40$ rule of thumb commonly used to identify acceptable latent trait indicators in exploratory factor analyses.

Isometric strength. Isometric strength tests measure the maximum force that a muscle can generate in a contraction that develops force, but the muscle does not shorten (Powers & Howley, 1990). Twenty-three of 24 isometric strength tests were acceptable; neck flexion was the exception (Table 3). The best indicators were low lift, $\lambda_{Avg} = .884$; shoulder extension, $\lambda_{Avg} = .828$; and torso/upper body flexion, $\lambda_{Avg} = .816$.

Isotonic strength. Isotonic strength tests involve contractions in which the muscle shortens against a fixed resistance. The shortening results in movement (Powers & Howley, 1990). All six tests were acceptable (Table 4). The best options were bench press, $\lambda_{Avg} = .856$; and shoulder press, $\lambda_{Avg} = .851$.

Dynamic strength. Dynamic strength tests required coordinated lifting actions involving multiple muscle groups. These tests were akin to Olympic weight lifts. The dynamic strength tests in this review were performed with an incremental lift machine. Stevenson, Bryant, Greenhorn, Deakin, and Smith (1995) described the lift dynamics. Both lift tests were acceptable (Table 5). The λ_{Avg} difference was too small to designate either test as the better option.

General strength. The isometric, isotonic, and dynamic strength latent traits were highly correlated, $.758 \leq r \leq .848$ (see Table 6). Exploratory factor analysis of the latent trait correlations produced a unidimensional model with the following factor loadings: Isoinertial Strength, $\lambda = .954$; Isometric Strength, $\lambda = .854$; and Dynamic Strength, $\lambda = .889$.

Muscular Endurance

Seven of nine ME tests were acceptable (Table 7); leg lifts and half-hold sit-ups were the exceptions. Dips, $\lambda_{Avg} = .761$; push-ups, $\lambda_{Avg} = .753$, pull-ups, $\lambda_{Avg} = .720$; and bent-arm hang $\lambda_{Avg} = .699$, were notably superior to the other ME tests, including sit-ups, $\lambda_{Avg} = .498$.

Muscular Power

All 12 MP tests were acceptable (Table 8). The best MP indicators were the 100-yd dash, $\lambda_{Avg} = .812$, and the 300-yd run, $\lambda_{Avg} = .786$, but those tests have been infrequently studied. If attention were limited to those tests that have been studied frequently ($k \geq 10$), the best tests were the 50-yd dash, $\lambda_{Avg} = .764$; the long jump, $\lambda_{Avg} = .734$; the vertical jump, $\lambda_{Avg} = .672$; and the medicine ball throw/shot put ($\lambda_{Avg} = .699$). The λ_{Avg} for each of these frequently studied tests fell within the 95% confidence intervals for the 100-yd dash and the 300-yd run, so all six tests were statistically equivalent. The λ_{Avg} for ergometer tests were substantially lower than those for dashes and jumps: arm ergometer, $\lambda_{Avg} = .559$; leg ergometer, $\lambda_{Avg} = .609$.

Cardiovascular Endurance

Eight of nine CE tests were λ_{Avg} acceptable; the step test was the exception (Table 9). Overlapping confidence intervals for the distance runs made it impossible to designate any best choice(s). The average factor loading for VO_{2max}, $\lambda_{Avg} = .707$, was notably weaker than that for any run test.

Latent Trait Correlations

Correlations between the physical ability latent traits were moderate, $.448 \leq r \leq .687$, except for a near-zero correlation of MS with CE ($r = .088$, see Table 10).

CFA Constraints

The CFA models estimated a factor loading for each test on a single factor. The models could have included a factor loading for each test on all four factors. However, the models fixed three possible factor loadings for each test at zero. These constraints on secondary factor loadings might have been inappropriate. Performance on some tests might be influenced by two or more physical abilities. The CFA analysis provided information that was used to evaluate constraint appropriateness. In particular, the output included estimates of what the secondary factor loadings would have been if they had been freely estimated.

Constraint appropriateness was evaluated by examining the 77 secondary factor loadings that had been estimated in three or more analyses. The average estimated secondary loading was -.020 across all 77 evaluated pairs. Only 4 of 77 pairs produced $|\lambda_{Est}| > .40$. A single outlier value accounted for the large average loading in each of those four cases.

Discussion

Standard practices correctly classify fitness tests in relation to general physical abilities. Using the standard classifications as the basis for CFA models, the fitness tests were acceptable ability indicators in 58 of 61 cases. The CFA models also provided enough information to evaluate the appropriateness of fixing secondary loadings at zero. The expected value of those constrained loadings was virtually zero. The expectations were not large enough to justify adding any secondary factor loadings given the risk of introducing post hoc model modifications based on chance findings (MacCallum, Roznowski, & Necowitz, 1992). Thus, 58 of 61 tests were acceptable indicators of their specified ability construct and were not related to any of the other ability constructs.

The physical ability constructs were correlated. The typical inter-trait correlation was moderately large. A near-zero correlation of MS with CE was the exception to this general trend. The latent trait correlations establish the potential for omitted variable bias (James, Mulaik, &

Brett, 1982). Earlier work demonstrated that bias can occur in causal models relating physical ability to physical task performance (Vickers, Hodgdon, & Beckett, 2009).

Having latent trait correlations in the measurement model was unusual. Past work has relied on orthogonal factor models. A model with correlated dimensions is consistent with the subjective impression that people differ in general fitness. A model with correlated dimensions also is more parsimonious. In the present case, six correlations between four latent traits have been substituted for the 183 secondary factor loadings that would be required for a four-dimensional orthogonal model for 61 ability tests. Model parsimony and plausibility both favor a correlated abilities model over an orthogonal abilities model.

The ability constructs represent performance capacities or physical proficiencies. These constructs should not be equated with specific physiological processes. The relatively modest factor loadings for laboratory tests of anaerobic and aerobic capacities support this view. If the CFA measurement models had been recast as causal models, the laboratory tests would have defined physiological constructs that caused performance differences. Had this been done, the physiological latent trait processes typically would have been identical to the laboratory test. The identity would occur because most studies included only one laboratory test for the relevant physiological capacities. Given the laboratory test–physiological process identity, the estimated causal effects of the physiological processes on performance would have been identical to the laboratory test factor loadings in the CFA measurement models. This factor loading interpretation would mean that anaerobic power accounts for 30% of the MP variance if the arm ergometer test is chosen, and 36% of the MP variance if the leg ergometer test is chosen. Laboratory-based aerobic capacity measures account for 50% of the CE performance variance.

The factor loadings have a simple practical interpretation. The factor loadings are the correlation of test scores with the latent traits. This correlation can be transformed to answer the question “How accurately will test scores identify individuals with above average ability?” A simple classification rule would predict that an individual with an above average test score was above average on ability. Using this rule, Rosenthal and Rubin’s (1982) binomial effect size display (BESD) converts correlations into the percentage of individuals in a sample who will be correctly classified using the stated rule. In the current context, BESD = 50 + (50* λ_{Avg}). The median BESD was 88% (range = 67%–99%). This figure is substantially higher than the 50% accuracy that would be expected if no test were given. With no information, accurate prediction would be expected in 50% of all cases. Thus, another interpretation is that the λ_{Avg} value for a test is the proportional reduction in error (PRE) associated with using that test instead of guessing (Hildebrand, Laing, & Rosenthal, 1977).

BESD and PRE provide a frame of reference for choosing between tests. Suppose Test A requires less time and equipment to administer than Test B. If Test B is less accurate, Test A is the clear choice. If Test B is more accurate than Test A, the choice becomes more complex. Test accuracy must be weighed against administrative simplicity. The accuracy difference will be too small to be important in many comparisons.

Even apparently large accuracy differences must be treated with caution. Some λ_{Avg} estimates are based on data from a few small samples. Those estimates will tend to have large 95% confidence intervals. A conservative treatment would consider this fact when comparing tests. Suppose λ_{Avg} for Test A is greater than λ_{Avg} for Test B. Tests A and B still could be regarded as equivalent if the 95% confidence interval for Test A included the Test B λ_{Avg} estimate. If the 95% confidence interval for Test A is broad, it can be appropriate to consider tests as equivalent even though the difference in their λ_{Avg} values is large. This issue is most likely to arise when considering tests that have been used infrequently in the past. Those tests are the ones that are likely to have wide confidence intervals. Additional study of promising alternatives that have been infrequently used in past research could be useful.

The results provide some general guidelines for test battery design. On the whole, the evidence supports the common practice of focusing on administrative simplicity. Usually, several tests have λ_{Avg} values that make them equivalent ability indicators for practical purposes. Administrative simplicity is a reasonable basis for choosing among those tests.

Test battery measurement precision can be increased by including multiple indicators for MS, MP, and ME, when possible. The tests in these three domains have moderate λ_{Avg} values. However, the tests load on the same factor because they are correlated. The sum of the standardized scores for two or more tests will estimate true ability more accurately than any single test (Nunnally & Bernstein, 1994). Choosing the tests with the highest λ_{Avg} values will maximize accuracy. Note that using multiple CE tests will have little value because the large λ_{Avg} values for run tests leave little room for improving the precision of a single test.

The limitations of this work must be considered to evaluate the results properly. The available evidence is skewed toward school-aged children and collegians in physical education classes. Tests have not been randomly paired within or across the ability domains. The analyses treat test administration differences (e.g., push-ups in 1 min or 2 min) as random variance sources. Constructs may not have been correctly interpreted. Lower body tests defined MP, and upper body tests defined ME. Perhaps both traits are narrow expressions of a general capacity for repetitive submaximal exertions. Some relevant data had to be omitted. Marsh (1993) produced a structural equation model that demonstrated invariance of physical ability latent traits across sex and age groups in a large sample. The assignments of tests to latent traits was sufficiently different from the assignments in this analysis to equate results from that study to the present findings. Finally, the lack of simple search terms to identify studies reporting inter-test correlations makes it almost certain that the literature search has overlooked some useful correlation matrices.

The ideal outcome would have been the identification of the best possible physical ability test battery. Instead, the evidence indicates that a number of equivalent test batteries can be constructed by defining sets of practically equivalent tests as the best choices within each physical ability domain. Equivalent test batteries then can be constructed by selecting one or more from each of the four “best test” sets. The failure to define the best possible test battery might be regarded as an outcome limitation, but guidance on how to construct equivalent test batteries may be a more useful outcome. This outcome provides the practitioner with flexibility in battery design coupled with confidence that his or her battery is optimum or close enough for practical applications.

Despite limitations, this review has produced several useful findings. The common treatment of MS, MP, ME, and CE as distinct ability constructs was supported. The results went beyond this simple affirmation by showing that the ability constructs are correlated in the general population. The analyses sharpened the interpretation of the ability constructs by highlighting the fact that these constructs represent performance capacities that should not be equated with specific physiological processes. Estimates of the effects of anaerobic and aerobic capacities on MP and CE performance were obtained as incidental modeling outcomes. The evidence supported the standard mapping of tests onto ability constructs and showed that tests are specific to a particular physical ability once the correlations between abilities are recognized. Test battery design has been facilitated by providing a set of λ_{Avg} values suitable for designing efficient, reliable fitness test batteries.

References

References marked with an asterisk indicate studies included in the meta-analysis.

Arbuckle, J. L., & Wothke, W. (1999). *Amos 4.0 user's guide*. Chicago: SmallWaters Corporation.

*Arnold, J. D., Rauschenberger, J. M., Soubel, W. G., & Guion, R. M. (1982). Validation and utility of a strength test for selecting steelworkers. *Journal of Applied Psychology, 67*, 588-604.

*Baumgartner, T. A., & Zuidema, M. A. (1972). Factor analysis of physical fitness tests. *Research Quarterly, 43*, 443-450.

*Beckett, M. B., & Hodgdon, J. A. (1987). *Lifting and carrying capacities relative to physical fitness measures* (NHRC Tech. Rep. No. 87-26). San Diego, CA: Naval Health Research Center.

*Bernauer, E. M., & Bonnano, J. (1975). Development of physical profiles for specific jobs. *Journal of Occupational Medicine, 17*, 27-33.

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

*Borchardt, J. W. (1968). A cluster analysis of static strength tests. *Research Quarterly, 39*, 258-264.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, West Sussex, United Kingdom: John Wiley and Sons.

*Burke, E. J. (1976). Validity of selected laboratory and field tests of physical working capacity. *Research Quarterly, 47*, 95-103.

*Carter, G. H., & Clarke, H. H. (1959). Oregon simplifications of the strength and physical fitness indices. *Research Quarterly, 30*, 3-10.

*Clarke, H. H. (1966). *Muscular strength and endurance in man*. Englewood Cliffs, NJ: Prentice-Hall.

*Clarke, H. H., & DeGutis, E. W. (1962). Comparison of skeletal age and various physical and motor factors with the pubescent development of 10, 13, and 16 year old boys. *Research Quarterly, 33*, 356-368.

*Costill, D. L., Miller, S. J., Myers, W. C., Kehoe, F. M., & Hoffman, W. M. (1968). Relationship among selected tests of explosive leg strength and power. *Research Quarterly, 39,* 785-787.

*Cozens, F. W. (1940). Strength tests as measures of general athletic ability in college men. *Research Quarterly, 11*, 45-52.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin, 105*, 317-327.

*Cureton, K. J., Boileau, R. A., Lohman, T. G., & Misner, J. E. (1977). Determinants of distance running performance in children: Analysis of a path model. *Research Quarterly, 48*, 270-279.

*Deason, J., Powers, S. K., Lawler, J., Ayers, D., & Stuart, M. K. (1991). Physiological correlates to 800 meter running performance. *Journal of Sports Medicine and Physical Fitness, 31*, 499-504.

*Dempsey, P. G., Ayoub, M. M., & Westfall, P. H. (1998). Evaluation of the ability of power to predict low frequency lifting capacity. *Ergonomics, 41*, 1222-1241.

*Disch, J., Frankiewicz, R., & Jackson, A. (1975). Construct validation of distance run tests. *Research Quarterly, 46*, 169-176.

*Falls, H. B., Ismail, A. H., & MacLeod, D. F. (1966). Estimation of maximum oxygen uptake in adults from AAHPER youth fitness test items. *Research Quarterly, 37*, 192-201.

*Farrell, P. A., Wilmore, J. H., Coyle, E. F., Billing, J. E., & Costill, D. L. (1979). Plasma lactate accumulation and distance running performance. *Medicine and Science in Sports, 11*, 338-344.

*Fay, L., Londeree, B. R., LaFontaine, T. P., & Volek, M. R. (1989). Physiological parameters related to distance running performance in female athletes. *Medicine and Science in Sports and Exercise, 21*, 319-324.

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

*Gutin, B., Fogle, R. K., & Stewart, K. (1976). Relationship among submaximal heart rate, aerobic power, and running performance in children. *Research Quarterly, 47*, 536-540.

*Hazard, A. A. (1982). *The effects of endurance training at 2,440m altitude on maximal oxygen uptake at altitude and sea level in young male and female middle distance runners.* Unpublished master's thesis, San Diego State University, San Diego, CA.

*Hetzler, R. K., DeRenne, C., Buxton, B. P., Ho, K. W., Chai, D. X., & Seichi, G. (1997). Effects of 12 weeks of strength training on anaerobic power in prepubescent male athletes. *Journal of Strength and Conditioning Research, 11*, 174-181.

Hildebrand, D. K., Laing, J. D., & Rosenthal, H. (1977). *Prediction analysis of cross-classifications*. NY: John Wiley & Sons.

*Hinton, E. A., & Rarick, L. (1940). The correlation of Rogers' test of physical capacity and the Cubberley and Cozens measurement of achievement in basketball. *Research Quarterly, 11*, 58-65.

*Hogan, J. (1985). Tests for success in diver training. *Journal of Applied Psychology, 70*, 219-224.

Hogan, J. (1991). Structure of physical performance in occupational tasks. *Journal of Applied Psychology, 76*, 495-507.

*Hutto, L. E. (1938). Measurement of the velocity factor and of athletic power in high school boys. *Research Quarterly, 9*, 109-128.

*Ismail, A. H., Christian, J. E., & Kessler, W. V. (1963). Body composition relative to motor aptitude for preadolescent boys. *Research Quarterly, 34*, 462-470.

*Jackson, A. S., & Osburn, H. G. (1983). *Validity of isometric strength tests for predicting performance in underground coal mining tasks*. Shell Oil Company.

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). *Causal analysis: Assumptions, models, and data*. Beverly Hills, CA: Sage Publications.

Joreskog, K. G., Sorbom, D., du Toit, S., & du Toit, M. (2000). *LISREL 8: New statistical features*. Chicago, IL: Scientific Software International.

*Joseph, J. J. (1990). The relationship of selected physical fitness test scores to the chronic complaints and ailments of adult males. *Journal of Sports Medicine and Physical Fitness, 7*, 83-94.

*Krahenbuhl, G. S., Pangrazi, R. P., Petersen, G. W., Burkett, L. N., & Schneider, M. J. (1978). Field testing of cardiorespiratory fitness in primary school children. *Medicine and Science in Sports, 10*, 208-213.

*Laughery, K. R., & Jackson, A. S. (1984). *Pre-employment physical test development for roustabout jobs on offshore production facilities*. Lafayette, LA: Kerr McGee Corporation.

*Liba, M. R. (1967). Factor analysis of strength variables. *Research Quarterly, 38*, 649-662.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*, 490-504.

*MacNaughton, L., Croft, R., Pennicott, J., & Long, T. (1990). The 5 and 15 minute runs as predictors of aerobic capacity in high school students. *Journal of Sports Medicine and Physical Fitness, 30*, 24-28.

*Manning, J. M., Dooly-Manning, C., & Perrin, D. H. (1988). Factor analysis of various anaerobic power tests. *Journal of Sports Medicine and Physical Fitness, 28*, 138-144.

Marsh, H. W. (1993). The multidimensional structure of physical fitness: Invariance over gender and age. *Research Quarterly for Exercise and Sport, 64*, 256-273.

*McCloy, E. (1935). Factor analysis methods in the measurement of physical abilities. *Research Quarterly*, *6* (Suppl.), 114-121.

*McCraw, L. W. (1949). A factor analysis of motor learning. *Research Quarterly, 20*, 316-335.

*McHone, V. L., Tompkin, G. W., & Davis, J. S. (1952). Short batteries of tests measuring physical efficiency for high school boys. *Research Quarterly, 23*, 82-93.

*Murphy, A. J., & Wilson, G. J. (1996). The assessment of human dynamic muscular function: A comparison of isoinertial and isokinetic tests. *Journal of Sports Medicine and Physical Fitness, 36*, 169-177.

*Myers, D. C., Gebhardt, D. L., Crump, C.E., & Fleishman, E. A. (1984). *Factor analysis of strength, cardiovascular endurance, flexibility, and body composition measures* (ARRO Technical Report 3077/R83-9). Bethesda, MD: Advanced Research Resources Organization.

*Myers, D. C., Gebhardt, D. L., Crump, C. E., & Fleishman, E. A. (1984). *Validation of the Military Entrance Physical Strength Capacity Test* (ARRO Tech. Rep. No. 610). Bethesda, MD: Advanced Research Resources Organization.

Nicks, D. C., & Fleishman, E. A. (1962). What do physical fitness tests measure? A review of factor analytic studies. *Educational and Psychological Measurement, 22*, 77-95.

*Noakes, T. D., Myburgh, K. H., & Schall, R. (1990). Peak treadmill running velocity during the VO2max test predicts running performance. *Journal of Sports Sciences, 8,* 35-45.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

*Padilla, S., Bourdin, M., Barthelemy, J. C., & Lacour, J. R. (1992). Physiological correlates of middle-distance running performance: A comparative study between men and women. *European Journal of Applied Physiology and Occupational Physiology, 65*, 561-566.

*Phillips, M. (1941). Study of a series of physical education tests by factor analysis. *Research Quarterly, 12*, 60-71.

*Ponthieux, N. A., & Barker, D. G. (1963). An analysis of the AAHPER Youth Fitness Test. *Research Quarterly, 34*, 525.

Powers, S. K., & Howley, E. T. (1990). *Exercise physiology: Theory and application to fitness and performance*. Dubuque, IA: Wm C. Brown Publishers.

*Rasch, P. J. (1974). Maximal oxygen intake as a predictor of performance in running events. *Journal of Sports Medicine, 14*, 32-39.

*Reilly, T., & Thomas, V. (1977). Effects of a programme of pre-season training on the fitness of soccer players. *Journal of Sports Medicine, 17*, 401-412.

*Robertson, D. W. (1982). *Development of an occupational strength test battery (STB)* (NPRDC Tech. Rep. 82-42). San Diego, CA: Navy Personnel Research and Development Center.

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74*, 166-169.

*Seashore, H. G. (1942). Some relationships of fine and gross motor abilities. *Research Quarterly, 13*, 259-274.

*Seymour, E. W. (1960). Follow-up study on simplification of the strength and physical fitness indexes. *Research Quarterly, 31*, 208-216.

*Sharp, D. S., Wright, J. E., Vogel, J. A., Patton, J. F., Daniels, W. L. Knapik, J., & Kowal, D. M. (1980). *Screening for physical capacity in the US Army: An analysis of measures predictive of strength and stamina* (USARIEM Rep. No. T 8/80). Natick, MA: U.S. Army Research Institute of Environmental Medicine.

*Sills, F. D. (1950). A factor analysis of somatotypes and their relationship to achievement in motor skills. *Research Quarterly, 21*, 424-437.

*Singh, M., Lee, S. W., Wheeler, G. D., Chahal, P., Oseen, M., & Couture, R. T. (1991). *Development of force mobile command army physical fitness evaluation and standards* *for field units* (Final Report). Edmonton, Alberta, Canada: Faculty of Physical Education and Recreation, University of Alberta.

*Sinnett, A. M., Berg, K., Latin, R. W., & Noble, J. M. (2001). The relationship between field tests of anaerobic power and 10-km run performance. *Journal of Strength and Conditioning Research, 15*, 405-412.

*Stanish, H. I., Wood, T. M., & Campagna, P. (1999). Prediction of performance on the RCMP Physical Ability Requirement Evaluation. *Journal of Occupational and Environmental Medicine*, *41*, 669-677.

Stevenson, J., Bryant, T., Greenhorn, D., Deakin, J., & Smith, T. (1995). Development of factor-score-based models to explain and predict maximal box-lifting performance. *Ergonomics, 38*, 292-302.

*Teves, M. A., Wright, J. E., & Vogel, J. A. (1985). *Performance on selected candidate screening test procedures before and after army basic and advanced individual training* (USARIEM Rep. No. T13/85). Frederick, MD: U.S. Army Medical Research and Development Command.

*Tharp, G. D., Newhouse, R. K., Uffelman, L., Thorland, W. G., & Johnson, G. O. (1985). Comparison of sprint and run times with performance on the Wingate Anaerobic Test. *Research Quarterly for Exercise and Sport, 56*, 73-76.

*Thorland, W. G., Johnson, G. O., Cisar, C. J., Housh, T. J., & Tharp, G. D. (1987). Strength and anaerobic responses of elite young female sprint and distance runners. *Medicine and Science in Sports and Exercise, 19*, 56-61.

*Tornvall, G. (1963). Assessment of physical capabilities: With special reference to the evaluation of maximal voluntary isometric muscle strength and maximal working capacity. *Acta Physiologica Scandinavica, 58*(Suppl. 201), 1-102.

Vickers, R. R., Jr. (2001a). *Running performance as an indicator of* VO_{2max}*: A replication of distance effects* (NHRC Tech. Rep. No. 01-24). San Diego, CA: Naval Health Research Center.

Vickers, R. R., Jr. (2001b). *Running performance as an indicator of* VO_{2max}*: Distance effects* (NHRC Tech. Rep. No. 01-20). San Diego, CA: Naval Health Research Center.

Vickers, R. R., Jr. (2003). *The measurement structure of strength* (NHRC Tech. Rep. No. 03-30). San Diego, CA: Naval Health Research Center.

Vickers, R. R., Jr., Hodgdon, J. A., & Beckett, M. B. (2009). *Physical ability-task performance models: Assessing the risk of omitted variable bias* (NHRC Tech. Rep. No. 09-04). San Diego, CA: Naval Health Research Center.

*Wiley, J. F., & Shaver, L. G. (1972). Prediction of maximum oxygen intake from running performance of untrained young men. *Research Quarterly, 43*, 89-93.

*Wright, J. E., Sharp, D. S., Vogel, J. A., & Patton, J. F. (1984). *Assessment of muscle strength and prediction of lifting capacity in U.S. Army personnel* (Technical Report). Natick, MA: U.S. Army Research Institute for Environmental Medicine.

*Zuidema, M. A., & Baumgartner, T. A. (1974). Second factor analysis study of physical fitness tests. *Research Quarterly, 45*, 247-256.

Table 1

Sample Descriptions

	k	*N*	No. of Test Scores
Adults			
Men	41	6,680	57,566
Women	18	3,200	26,784
Men and women	3	468	17,699
Children			
Boys	1	20	60
Girls	2	118	854
Boys and girls	14	2,042	14,617
Totals			
Male	42	6,700	57,626
Female	20	3,318	27,638
Adult	62	10,348	102,049
Child	17	2,180	15,331
No information	5	743	5,724
Grand total	84	13,271	123,104

Note. Cumulative values for age and gender groups do not equal the total because they do not include those samples for which no demographic information was available. The table omits Blakly et al.'s (1994) sample of *N* = 13,000 men and women who contributed 52,000 test scores so that one exceptional sample did not inflate the totals.

Table 2

Data Distribution by Ability Construct

	k	*N*	No. of Test Scores
Isometric strength	44	6,808	30,440
Isotonic strength	7	1,455	3,068
Dynamic strength	9	1,315	6,455
Muscular endurance	36	10,112	32,024
Muscular power	37	7,390	28,680
Cardio endurance	30	2,747	7,136
Total	84	13,271	107,803

Note. The tabled values omit Blakly et al.'s (1994) sample of $N = 13{,}000$ men and women who contributed 52,000 test scores so that one exceptional sample did not inflate the totals.

Table 3

Isometric Strength Test Results

				95% CI Bounds					
Test	*k*	λ_{Avg}	*SE*	Lower	Upper	*Q*	Sig	*z*	Sig
Low lift	9	.884	.031	.826	.942	6.49	.592	15.49	.000
Shoulder extension	3	.828	.033	.731	.924	1.53	.465	12.94	.000
Torso/upper body flexion	4	.816	.052	.692	.939	3.09	.377	7.93	.000
Back dynamometer	4	.788	.095	.563	1.012	2.42	.490	4.06	.000
Hip flexion	5	.782	.046	.684	.881	2.27	.685	8.26	.000
Shoulder flexion	4	.776	.024	.719	.834	2.90	.408	15.48	.000
Medium lift	9	.763	.031	.706	.820	7.56	.477	11.84	.000
Elbow flexion	7	.762	.042	.681	.843	4.68	.586	8.71	.000
Back lift	12	.737	.043	.660	.814	15.59	.157	7.84	.000
Knee extension	11	.723	.040	.651	.796	9.48	.487	8.10	.000
Arm dynamometer	5	.723	.128	.449	.997	2.83	.587	2.52	.006
Leg lift	11	.716	.030	.661	.771	13.59	.193	10.38	.000
Arm lift	13	.692	.024	.649	.736	13.09	.362	12.05	.000
Trunk flexion	11	.690	.032	.632	.747	13.34	.205	9.16	.000

(continued)

Table 3 (continued)

Isometric Strength Test Results

				95% CI Bounds					
Test	*k*	λ_{Avg}	*SE*	Lower	Upper	*Q*	Sig	*z*	Sig
Arm pull	13	.684	.026	.637	.730	11.74	.467	10.87	.000
Ankle plantarflexion	7	.675	.050	.577	.772	6.08	.414	5.47	.000
Trunk extension	16	.667	.020	.631	.703	15.16	.440	13.10	.000
Elbow extension	4	.666	.053	.541	.791	3.68	.298	4.99	.000
Handgrip	35	.652	.021	.616	.688	33.68	.483	11.94	.000
Knee flexion	6	.648	.063	.521	.775	3.91	.562	3.93	.000
Hip extension	10	.623	.064	.506	.740	6.20	.719	3.50	.000
Neck extension	3	.599	.053	.444	.754	1.94	.379	3.75	.000
Ankle dorsiflexion	5	.556	.054	.440	.671	4.40	.355	2.88	.002
Neck flexion	3	.492	.099	.203	.780	2.57	.276	.93	.177

Note. *k* is the number of samples that provided results for the test. The table includes all isometric strength tests for which $k \geq 3$. λ_{Avg} is the weighted average factor loading from the random effects analysis. CI is confidence interval. *Q* is a measure of dispersion of the random effects estimates. This statistic has an asymptotic χ^2 distribution with $k - 1$ *df*. *z* is a test of the hypothesis that $\lambda_{Avg} > .40$. *SE* = standard error.

Table 4

Isotonic Strength Test Results

				95% CI Bounds					
Test	*k*	λ_{Avg}	*SE*	Lower	Upper	*Q*	Sig	*z*	Sig
Bench press	8	.856	.038	.785	.928	7.08	.421	12.09	.000
Shoulder press	7	.851	.025	.802	.900	2.76	.838	17.83	.000
Lat pull-down/trapezius	7	.797	.028	.743	.852	5.01	.542	14.20	.000
Arm curl	8	.750	.036	.682	.818	7.98	.334	9.79	.000
Knee ext	4	.607	.056	.475	.740	3.40	.334	3.67	.000
Leg extension	9	.603	.032	.543	.663	8.78	.362	6.27	.000

Note. k is the number of samples that provided results for the test. The table includes all isometric strength tests for which k ≥ 3. λ_{Avg} is the weighted average factor loading from the random effects analysis. CI is confidence interval. *Q* is a measure of dispersion of the random effects estimates. This statistic has an asymptotic χ^2 distribution with *k* – 1 *df*. *z* is a test of the hypothesis that $\lambda_{Avg} > .40$. *SE* = standard error.

Table 5

Dynamic Strength Test Results

				95% CI Bounds					
Test	*k*	λ_{Avg}	*SE*	Lower	Upper	*Q*	Sig	*z*	Sig
ILM high	7	.928	.021	.887	.969	3.45	.751	25.02	.000
ILM low	7	.856	.047	.766	.946	7.52	.275	9.80	.000

Note. ILM high = incremental lift machine lift to 180 cm; ILM low = incremental lift machine lift to 152 cm. *k* is the number of samples that provided results for the test. The table includes all isometric strength tests for which $k \geq 3$. λ_{Avg} is the weighted average factor loading from the random effects analysis. CI is confidence interval. *Q* is a measure of dispersion of the random effects estimates. This statistic has an asymptotic χ^2 distribution with $k - 1$ *df*. *z* is a test of the hypothesis that $\lambda_{Avg} > .40$.

Table 6

Correlations of Modality-Specific Strength Factors

Factor	1	2	3
1. Isotonic	1.000		
2. Isometric	.815 (*k* = 6)	1.000	
3. Dynamic	.848 (*k* = 3)	.758 (*k* = 7)	1.000

Note. Table entries are the pooled correlations between the general ability factors. The *k* values are the number of samples that provided estimates of each correlation. Isotonic = isotonic strength.

Table 7

Muscular Endurance Test Results

				95% CI Bounds					
Test	*k*	λ_{Avg}	*SE*	Lower	Upper	*Q*	Sig	z^a	Sig
Dips	7	.761	.051	.662	.861	5.16	.523	7.06	.000
Push-up	20	.753	.038	.687	.818	18.40	.496	9.32	.000
Pull-up	30	.720	.030	.669	.770	25.28	.664	10.73	.000
Bent-arm hang	11	.699	.045	.617	.781	8.29	.601	6.63	.000
Endurance	14	.549	.067	.430	.667	14.21	.359	2.23	.013
Sit-up	27	.498	.023	.459	.538	21.57	.712	4.25	.000
Squat	10	.474	.038	.404	.544	7.53	.582	1.95	.026
Leg lift/raise	8	.421	.057	.313	.529	7.42	.387	.37	.356
Half-hold sit-up	6	.363	.030	.302	.424	4.38	.496	-1.22	.888

Note. *k* is the number of samples that provided results for the test. The table includes all isometric strength tests for which $k \geq 3$. λ_{Avg} is the weighted average factor loading from the random effects analysis. CI is confidence interval. *Q* is a measure of dispersion of the random effects estimates. This statistic has an asymptotic χ^2 distribution with $k - 1$ *df*. *z* is a test of the hypothesis that $\lambda_{Avg} > .40$. *SE* = standard error.

Table 8

Muscular Power Test Results

				95% CI Bounds					
Test	*k*	λ_{Avg}	*SE*	Lower	Upper	*Q*	Sig	*z*	Sig
100-yd dash	4	.812	.070	.648	.976	.87	.833	5.92	.000
300-yd run	4	.786	.077	.605	.966	3.24	.356	5.04	.000
10-yd dash	2	.782	.055	.434	1.130	.74	.389	6.94	.000
50-yd dash[a]	22	.764	.037	.700	.828	24.06	.290	9.84	.000
Shuttle run	8	.746	.060	.633	.860	13.62	.058	5.76	.000
Long jump	30	.734	.029	.685	.783	29.94	.417	11.62	.000
600-yd run	7	.705	.050	.608	.801	7.76	.256	6.14	.000
Vertical jump	25	.672	.026	.628	.717	21.09	.633	10.47	.000
Medicine ball/shot put	10	.664	.072	.531	.797	10.97	.278	3.64	.000
40-yd dash	5	.653	.151	.330	.975	4.53	.339	1.67	.048
Leg ergometer	8	.609	.068	.480	.737	6.09	.530	3.07	.001
Arm ergometer	6	.559	.053	.453	.666	4.86	.433	3.02	.001

Note. k is the number of samples that provided results for the test. The table includes all isometric strength tests for which $k \geq 3$. λ_{Avg} is the weighted average factor loading from the random effects analysis. CI is confidence interval. *Q* is a measure of dispersion of the random effects estimates. This statistic has an asymptotic χ^2 distribution with $k - 1$ *df*. *z* is a test of the hypothesis that $\lambda_{Avg} > .40$. *SE* = standard error.

[a] Includes one 60-yd dash.

Table 9

Cardiovascular Endurance Test Results

				95% CI Bounds					
Test	*k*	$\lambda_{\text{Avg}\,g}$	*SE*	Lower	Upper	*Q*	Sig	z^a	Sig
2-mi run	6	.908	.063	.781	1.034	.61	.987	8.10	.000
1-mi run	10	.891	.047	.804	.978	8.98	.439	10.36	.000
880-yd run	4	.889	.044	.785	.993	.67	.881	11.03	.000
3-mi run[a]	4	.886	.092	.670	1.102	.68	.877	5.30	.000
1.5-mi run	5	.880	.051	.772	.988	3.62	.460	9.50	.000
12-min run	11	.821	.038	.752	.891	8.54	.576	10.95	.000
1- to 1.2-km run	5	.792	.063	.658	.926	2.41	.660	6.24	.000
$V_{O2max}{}^b$	20	.707	.063	.598	.817	11.74	.896	4.85	.000
Step test	5	.362	.044	.268	.457	4.04	.401	-.85	.801

Note. Runs >10 km have been omitted from the table because it is unlikely that those distances would be considered for inclusion in fitness tests. *k* is the number of samples that provided results for the test. The table includes all isometric strength tests for which $k \geq 3$. λ_{Avg} is the weighted average factor loading from the random effects analysis. CI is confidence interval. Q is a measure of dispersion of the random effects estimates. This statistic has an asymptotic χ^2 distribution with $k - 1$ *df*. *z* is a test of the hypothesis that $\lambda_{Avg} > .40$. *SE* = standard error.

[a] Includes one 5-km run.

[b] Laboratory measurement of maximal oxygen uptake.

Table 10

Intercorrelations of Ability Factors

Ability Factor	Isometric MS	MP	ME	CE
Isometric MS	1.000			
MP	.572 (*k* = 11)	1.000		
ME	.448 (*k* = 18)	.687 (*k* = 24)	1.000	
CE	.088 (*k* = 4)	.504 (*k* = 13)	.595 (*k* = 11)	1.000

Note. Table entries are the pooled correlations between the general ability factors. The *k* values are the number of samples that provided estimates of each correlation. Isometric MS = isometric muscular strength; MP = muscular power; ME = muscular endurance; CE = cardiovascular endurance.

APPENDIX A

Muscle Strength Measurement Model

Should muscular strength be represented as a single general construct or is it more appropriate to represent muscular strength as a set of correlated dimensions representing different measurement methods? To answer this question, a unidimensional strength model was compared with a multidimensional model in 10 data sets that included ≥2 strength tests for ≥2 measurement methods. All of the strength tests loaded on a single factor in the unidimensional model. Tests loaded on isometric, isotonic, or dynamic strength factors, as appropriate for the test, in the multidimensional models. The Singh et al. (1991) model included only the right side measurement for each bilateral exercise.

The multidimensional goodness of fit was significantly better in 9 of the 10 analyses (Table A), so the cumulative improvement in fit was statistically significant ($\chi^2 = 394.54$, 16 *df*, $p < .001$). The consistent trend was more important than the significance in any given sample or the cumulative significance. The root mean square error of approximation, non-normed fit index, and standardized root mean residual, all of which are widely used goodness-of-fit indices indicated modest gains in absolute fit.

Table A

Comparison of Unidimensional With Multidimensional Strength Models

Model	# Dim	χ^2	Sig	RMSEA	NNFI	$\Delta\chi^2$	*df*	Δ NNFI
Beckett & Hodgdon								
Women	1	56.85	.000	.130	.794			
	3	55.00	.001	.139	.752	1.85	3	
Model comparison						1.85	3	-.042
Men	1	112.24	.000	.187	.817			
	3	80.21	.000	.155	.861			
Model comparison						32.03	3	.044
Marcinik studies								
Orlando	1	190.76	.000	.146	.817			
	3	113.55	.000	.106	.892	77.21	3	
Model comparison						77.21	3	.075
Shipboard	1	73.99	.000	.176	.837			
	2	63.46	.000	.160	.847			
Model comparison						10.53	1	.010
Sparten	1	102.25	.000	.150	.864			
	2	51.70	.003	.089	.943			
Model comparison						51.95	1	.089

(continued)

Table A (continued)

Comparison of Unidimensional With Multidimensional Strength Models

Model	# Dim	χ^2	Sig	RMSEA	NNFI	$\Delta \chi^2$	*df*	Δ NNFI
Myers et al.								
Men	1	76.34	.000	.273	.852			
	2	2.02	.156	.045	.996			
Model comparison						74.32	1	.144
Women	1	90.83	.000	.298	.744			
	2	.01	.912	.000	1.01			
Model comparison						89.82	1	.257
Singh et al.	1	503.28	.000	.200	.504			
	2	497.16	.000	.200	.501	6.12	1	
Model comparison						6.12	1	-.003
Teves et al.								
Men	1	34.79	.000	.262	.772			
	2	4.33	.363	.024	.997			
Model comparison						30.46	1	.225
Women	1	21.08	.001	.167	.840			
	2	.83	.934	.000	1.04			
Model comparison						20.25	1	.164

Note. Arbuckle and Wothke (1999, pp. 395-416) provide definitions of the root mean square error of approximation (RMSEA) and the non-normed fit index (NNFI). The $\Delta \chi^2$ column indicates the improvement in fit obtained by moving from the unidimensional model to the multidimensional model.

APPENDIX B

Evaluation of Meta-Analysis Parameter Weights

Accurate standard errors were essential for the planned meta-analyses. Accuracy was critical because these error estimates provided the basis for weighting the loadings when computing the pooled factor loadings (Borenstein, Hedges, Higgins, & Rothstein, 2009). Correct weighting was essential for valid analytical results.

Analyzing correlation matrices can distort standard error estimates (Joreskog, Sorbom, du Toit, & du Toit, 2000, Appendix C, pp. 209–214). Cudeck (1989) demonstrated two minimum requirements for obtaining accurate standard errors when correlation matrices are analyzed. All model parameters must be freely estimated and the reproduced correlation matrix must have ones on the diagonal. Every CFA model in this study satisfied the first requirement. The second condition held except in the data from Falls, Ismail, and MacLeod (1966), so that sample was dropped from the meta-analysis.

CFA models based on covariance matrices provided additional reason to accept the standard errors derived from analyses of correlation matrices. Standard deviations for the fitness test scores had been reported in some studies. Covariance matrices could be constructed for those studies by combining the standard deviations with the correlation matrices.

Covariance-based CFA models were evaluated for three studies. Two findings from those analyses supported the accuracy of the correlation models. First, the completely standardized factor loadings were identical to the corresponding loadings from the correlation analyses. Second, the *t* values for the corresponding factor loadings were identical in the two analyses. These results suggested that the CFA models were invariant under the transformation from covariance matrices to correlation matrices. Invariance under transformation is a third condition that is associated with accurate standard error estimations in correlation matrix analyses (Joreskog et al., 2000).

REPORT DOCUMENTATION PAGE

1. REPORT DATE (DD MM YY)
03 02 11

2. REPORT TYPE
Technical Report

3. DATES COVERED (from – to)
Jul 2007–Dec 2010

4. TITLE
Construct Validity of Physical Fitness Tests

5a. Contract Number:
5b. Grant Number:
5c. Program Element Number:
5d. Project Number:
5e. Task Number:
5f. Work Unit Number: 60704

6. AUTHORS
Ross R. Vickers, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Commanding Officer
Naval Health Research Center
140 Sylvester Rd
San Diego, CA 92106-3521

8. PERFORMING ORGANIZATION REPORT NUMBER
11-52

8. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES)
Commanding Officer
Naval Medical Research Center
503 Robert Grant Ave
Silver Spring, MD 20910-7500

Chief, Bureau of Medicine and Surgery
(MED 00), Navy Dept
2300 E Street NW
Washington, DC 20372-5300

10. SPONSOR/MONITOR'S ACRONYM(S)
NMRC/BUMED

11. SPONSOR/MONITOR'S REPORT NUMBER(s)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The process of defining physical fitness test batteries typically relies on qualitative evaluations of individual tests. Starting from the existing consensus regarding the mapping of physical fitness tests onto physical ability constructs, analyses were carried out to develop quantitative test validity indices for use in test battery design. The validity indices were averaged factor loadings from confirmatory factor analysis (CFA) of the inter-test correlation matrices from 85 independent samples. The CFA included latent traits representing muscular strength, muscular power, muscular endurance, and cardiorespiratory endurance. The averaged factor loadings came from random effects analysis of the factor loadings from the 85 measurement models. The results confirmed the accepted assignment of fitness tests to categories representing the four physical ability constructs. The average factor loading varied from test to test within each category, but the inter-test variation generally was small relative to the standard errors of the individual loading estimates. The modest validity differences leave considerable freedom to use additional criteria, such as ease of administration, time requirements, and face validity from the perspective of the test population, when designing physical fitness test batteries.

15. SUBJECT TERMS
physical fitness test, physical abilities, test validity, meta-analysis

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**18a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Commanding Officer
UNCL	UNCL	UNCL	UNCL	43	**18b. TELEPHONE NUMBER (INCLUDING AREA CODE)** COMM/DSN: (619) 553-8429

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18