NAVAL HEALTH RESEARCH CENTER

PHYSICAL TEST VALIDATION FOR JOB SELECTION

CHAPTER 5

J. A. Hodgdon

A. S. Jackson

Report No. 01-12

Approved for public release; distribution unlimited.

NAVAL HEALTH RESEARCH CENTER
P. O. BOX 85122
SAN DIEGO, CA 92186-5122

BUREAU OF MEDICINE AND SURGERY (MED-02)
2300 E ST. NW
WASHINGTON , DC 20372-5300

PHYSICAL TEST VALIDATION FOR JOB SELECTION

Chapter 5

James A. Hodgdon, Ph.D [1]

Andrew S. Jackson, P.E.D [2]

[1] Naval Health Research Center
P. O/ Box 85122
San Diego, California 92186-5122

and

[2] Department of Health and Human Performance
University of Houston
Houstin, Texas

Published in: Constable & Palmer (Eds.), *The Process of Physical Fitness Standards Development*, Human Systems IAC SOAR, 2000.

Report No. 01-12, supported by the Navy Medical Research and Development Command, under research work unit 63706N M0096,002-6716. The views expressed in this article are those of the authors and do not 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 unlimited.

Chapter 5

Physical Test Validation for Job Selection

James A. Hodgdon, Ph.D.
Human Performance Center
Naval Health Research Center
San Diego, CA

Andrew S. Jackson, P.E.D.
Department of Health and Human Performance
University of Houston
Houston, TX

Abstract

This chapter examines the issues related to physical test validation for job selection. The chapter is divided into three major sections. The first examines issues and accepted methods of test validation. The focus is on the interpretation of the Equal Employment Opportunity Commission (EEOC) guidelines (EEOC, 1978) as they relate to test validation. The sanctioned validation methods are content validity, criterion-related validity, and construct validity. The measurement theory used to evaluate the quality of employment tests is based on the American Psychological Association standards for validating educational and psychological tests (A.P.A, 1985; A.P.A., 1987). A major difference in physical test validation is the use of physiological rather then psychological tests. The second section of the chapter examines the differences between physiological and psychological test validation. The goal of physiological validation is to define the physiological capacity needed by a worker to perform the work demanded by the task. Principal features of the physiological validation approach are the use of a physiological metric to quantify test performance and the interpretation of validity results with relevant physiological research and theory. The final section of the chapter reviews published employment validation research on physical tests.

Employment Selection Tests

The principal guidance for the design and implementation of selection tests for employment is the Uniform Guidelines on Employee Selection Procedures issued by the Equal Employment Opportunity Commission in 1978 (EEOC, 1978). These guidelines state that a selection procedure has "adverse impact" if the selection rate for any group is less than 80 percent for the group with the highest selection rate. Selection procedures that have adverse impact are considered discriminatory unless they can be justified. A selection procedure that has adverse impact can be justified if—

1. The tests or measures are derived from a job analysis
2. The tests or measures are indicators of critical or important job duties, work behaviors, or work outcomes
3. The tests or measures have been shown to be valid indicators of such duties, behaviors, or outcomes.

This existence of a procedure for justifying selection tests is critical in the area of selection based on physical abilities. There are well-recognized differences in physical abilities between genders (McArdle, Katch, & Katch, 1996), and the development of a physical abilities selection test for physically demanding jobs runs a great risk of having adverse impact across gender.

The nature of job analyses, identification of critical or important job duties, and nature of physical selection tests are discussed in other sections. This section considers issues surrounding the demonstration of the validity of selection tests or measures. As in other sections of this report, the emphasis is on selection based on physical ability.

Validity of Selection Tests

The extent to which a test or set of tests measures what it is meant to measure is called the validity of the test. For the purposes of this chapter, validity is the accuracy with which selection test(s) measure important work behaviors (Jackson, 1994). The Uniform Guidelines recognize three types of validity with respect to selection test development: content validity, criterion-related validity, and construct validity.

Content Validity—That a test has content validity means that the test items reflect important elements of the job. The job and test content are linked. Most content-valid test items are, in fact, job samples or simulations of job tasks. Theoretically, for the test as a whole to be content valid, the test items must sample all critical or important duties, work behaviors, or work outcomes. For example, if a job has two critical, physically demanding tasks, one involving repeated lifting to a fixed height and one involving carrying materials a long distance, both tasks need to be simulated in the content-valid selection test. Such job sample tasks are usually scored as to whether the applicant can or cannot perform the task. Additionally, for jobs that have time constraints, such as emergency service tasks, there may be time limits imposed for task completion. Successful completion of the tasks qualifies one for the job. Content-valid tests are the most defensible tests because they are the most direct indicators of job performance capability. The closer the simulation is to the actual job task, the more defensible it is as a selection test.

Criterion-Related Validity—A test is said to have criterion-related validity when the test items are shown to be estimators or predictors of critical or important duties, work behaviors, or work outcomes. Criterion-related validity is usually expressed as a correlation coefficient between test performance (the predictor) and performance of an important or critical job element or behavior (the criterion). The criterion job element can be any of a number of job behaviors including work-task performance, injury rates on the job, absenteeism, or peer or supervisor ratings. Criterion-related selection tests are not, by definition, direct indicators of the ability to perform a job or job task. They rely on a secondary relationship between the criterion task and the predictor test.

Two types of criterion-related validity can be distinguished. A test is said to have concurrent validity whenever the test is used to predict a current capability. An example is use of a bench press 1-repetition maximum (1RM) is used to predict an applicant's current ability to lift a 50-kg box to elbow height. If the test is used to predict some future event, it is said to have predictive validity. An example would be the use of the time to complete a 1-mile run as an indicator of future success in a Military training program.

Correlational studies are carried out to demonstrate criterion-related validity. Critical or important job behaviors are determined during the job analysis. The nature of the critical job behaviors usually suggests the nature of the selection test to be employed. If a critical task requires lifting, for example, then selection tests that measure strength would be appropriate. If the critical job task requires prolonged activity, then a test related to endurance, such as a run for time, might be appropriate. Once candidate tests have been chosen, the tests are administered to a sample of workers or another suitable sample. Their performance on the identified critical job tasks (or other criterion measures) is also measured. The strength of the associations between performance on the selection tests and performance on the critical job behaviors is expressed as the correlation coefficient, which is a measure of the amount of common variance accounted for by two measures. If the correlation coefficient between a selection test performance and performance on a critical job behavior is suitably high, the selection test may be used. It should be noted that there is no standard for the minimum acceptable correlation coefficient between a selection test and job behavior. Statistical significance is not always a good indicator because with large sample sizes, a correlation that explains only a small part of the variance can be significant. That which is possible or practical may drive the selection of an acceptable level of correlation. As a benchmark, one might note that a correlation coefficient of 0.707 indicates that 50 percent of the common variance in the relationship has been explained, but this is difficult to use as a criterion because many things can affect the size of a correlation coefficient. For example, the size of correlation is influenced substantially by the variability of the sample tested. It is also possible to have a high correlation but considerable errors in prediction (Altman & Bland, 1983; Altman & Bland, 1986). This subject is covered in more detail in another section of this chapter.

The scoring of criterion-related tests is based on the achievement of critical performance levels on the selection test(s). These critical performance levels can be quite difficult to define. Usually, they are derived from a mathematical function relating the predictor and criterion performances. The value of the performance on the selection test that is associated mathematically with a critical level of performance on the important job task is used as the cut off score or cut-score on the selection test. This critical level of job performance needs to be identified in the job analysis. This subject is covered in more detail in another section of this chapter.

Even in the simplest case, when a single critical task and critical level of performance, and a single predictor measure are identified, it can be difficult to set a critical level of performance. This is because the relationship between performance on the selection test and performance on the criterion task is not perfect. As an example, Figure 5.1 shows the relationship between the maximum weight box that can be lifted to elbow height (the work tasks) and 1RM for arm-curl (the criterion test). As one can see, arm-curl 1RM and maximum box weight appear to be strongly related. The correlation coefficient for this relationship is 0.875. Furthermore, the relationship between the variables appears to be a straight line, as suggested by the diagonal line crossing the figure. This line represents the linear regression of maximal box lift weight with arm-curl 1RM. However, the points are scattered about the line. If the critical task for a particular job involved lifting a 50-kg box to elbow height (the value indicated by the horizontal line), the mean arm-curl value associated with this box weight is 23.4 kg (the solid vertical line). This, ideally, would be the critical arm-

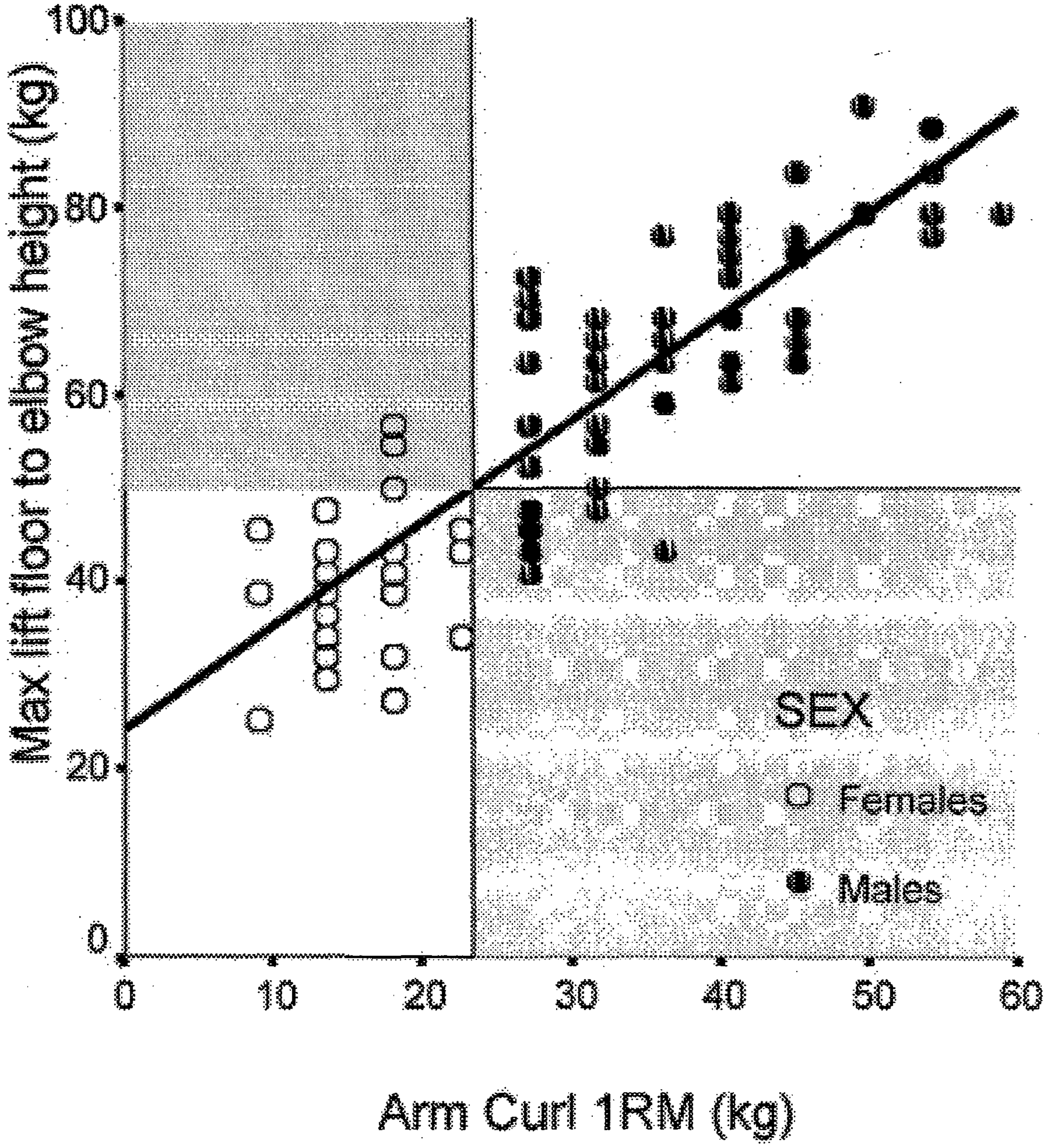

Figure 5.1 Maximum box weight lifted to elbow height as a function of arm curl 1RM

curl 1RM value that we would pick if arm-curl 1RM were the selection task for this job. However, it is clear by inspection of Figure 5.1 that some individuals who lifted less than 23.4 kg on the arm-curl could lift a 50-kg box. These individuals are called "false negatives" because they failed the test (and are not selected) but can perform the work task. In Figure 5.1, the false negatives appear in the upper left quadrant formed by the horizontal and vertical lines within the figure. Similarly, some individuals who lifted more than 23.4 kg could not lift a 50-kg box. These individuals are known as "false positives" because they passed the test (and were selected), but cannot perform the work task. In Figure 5.1, these individuals appear in the lower right quadrant. The Uniform Guidelines allow the exercise of a certain amount of judgment in setting cut-scores. However, one needs to have a defensible rationale. These issues are examined in more detail in the physiological validation section of this chapter.

Construct Validity—Construct validity is the most indirect and theory-driven method of establishing validity. Construct validity exists when selection tests are related to a general trait or set of characteristics (the construct) that is associated with successful accomplishment of important or critical job behaviors. The establishment of construct validity requires that employers show that a construct (a general trait or set of characteristics) is required for satisfactory job performance, and that the selection test or tests measure this same construct.

Constructs are often developed using the statistical technique of factor analysis (Rummel, 1970). In factor analysis, a number of correlated variables are reduced to a smaller number of dimensions or factors. Within the factor, each of the included variables has a coefficient or "loading," a numerical value indicating the strength of association of that variable with the factor. The greater the loading, the greater the association between the variable and the factor. The factor is defined mathematically as the sum of the factor variable values, each multiplied by its loading. The variables with the greatest loadings drive the theoretical interpretation of the factor.

Construct validity can be established in three ways—

1. Performances on job behaviors can be analyzed to determine dimensions within the job. Scores on selection tests can then be shown to be correlated with the job dimensions.
2. Scores on selection tests can be factor analyzed, and dimensions within the selection tests identified. A number of examples of such analyses can be found in the literature (Fleishman, 1964; Hogan, 1991a; Meyers, Gebhardt, Crump, & Fleishman, 1984)
3. Factor scores from the dimensions of the selection tests can be shown to be correlated to performance on important job behaviors. Both potential selection test items and performance on important job behaviors can be factor analyzed. A validity study can then be carried out to analyze the associations between the selection factors and the job factors.

These options are indicated schematically in Figure 5.2.

Figure 5.2 is an oversimplified version of the actual situation. Often, more than one construct is present in the job behaviors. For example, strength and endurance may be required for job success. In such a case, many more relationships must be worked out in the validity study.

The conduct of a study to demonstrate construct validity is similar to that for criterion-related validity except that instead of a one-to-one mapping of performance on a selection test to perform-

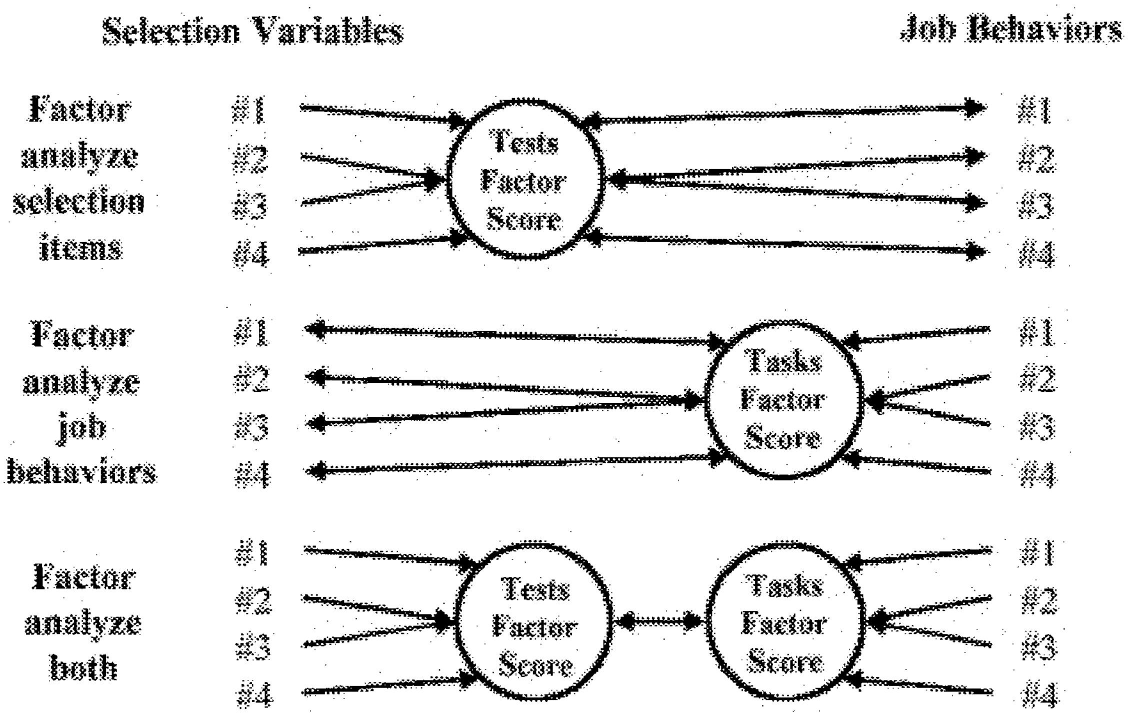

Figure 5.2 Three experimental designs for construct validity studies. Single-ended arrows indicate variables included in the factor analysis. Double-ended arrows indicate correlations to be measured.

ance on a job behavior, several selection-test items are measured that are used to calculate factor scores to represent the selection constructs being measured, and/or several job behaviors are measured to calculate factor scores to represent the job constructs being measured. It is these factor scores that are used in the correlational analysis. Construct-validity relationships are often difficult to demonstrate because of the need to identify the factor structures in the job and selection tests and then establish associations between or among them. Given these difficulties, many employers choose to use the measures of underlying constructs directly as elements of criterion-related validity studies.

Requirements for Validity Studies

The Uniform Guidelines provide general and technical standards for validity studies. Among the general standards are the following—

- In addition to specifying the three types of studies (content, criterion-related, and construct-validity), the guidelines require the studies to be consistent with applicable professional standards for such research, accurate and free from bias.
- The validity studies should be documented.
- The employer must be prepared to justify the method used to implement the selection tests. If use of a test has greater adverse impact when used as a ranking device than if it were

implemented as a simple pass/fail, then the employer must provide sufficient evidence of the validity and utility to support use of the test to rank-order participants.

- Selection procedures may be developed for higher level jobs in cases where most of the entry-level applicants will progress to those higher level jobs.
- An employer may continue to use selection procedures for which there is not yet full validity evidence as long as the employer has evidence of the substantial validity of the procedures and will conduct, when technically feasible, a study to produce the additional evidence required.
- Employers may also use validity studies conducted by others when it can be shown that the validity studies were conducted properly and that the jobs perform substantially the same major work behaviors for the employer as for those who conducted the study.
- Employers, labor organizations, and employment agencies are encouraged to work together and cooperate in validity studies.
- Finally, under no circumstances will the general reputation of a test or other selection procedures or casual reports of its validity be accepted in lieu of evidence of validity.

The minimum technical standards called for in the guidelines of all tests are that validity studies should be based on review of information about the job (a job analysis). The technical standards differ somewhat for the type of validation study. Tables 5.1, 5.2, and 5.3 summarize these standards by validation method.

Table 5.1 EEOC Technical standards Guidelines for the criterion validation method

Technical Standard for Criterion-Related Validation Studies
1. The study must be technically feasible. It must be possible to get an adequate sample size to provide a scientifically sound result. However, an employer is not required to hire or promote individuals in order to be able to conduct a criterion-related study.
2. Whether the study is to be concurrent or predictive, the sample subjects should be representative of the individuals who might reasonably be expected to fill the positions being studied.
3. In general, the guidelines indicate the finding of a significance level $P \leq 0.05$ to be acceptable.
4. However, users should evaluate each selection procedure to assure that it is appropriate for operational use. In general, the greater the magnitude of the correlations found between the job behaviors and the tests, and the greater the number of job behaviors predicted by a particular test, the more appropriate it is for implementation. Selection procedures derived from studies with large sample sizes and low correlations, and sole reliance on a selection instrument that is related to only one of many critical job behaviors will be subject to close review.
5. Users must avoid use of techniques that can lead to inflated validities for selection procedures. Examples include reliance on a few selection procedures or criteria when many were studied, and use of the statistics from one sample when they may not have held up well on cross-validation. The Guidelines recommend large samples and use of cross-validation.
6. The Guidelines call for the maintenance of "fairness" in selection procedures. Essentially, unfairness results when members of one group characteristically obtain lower scores on a selection procedure than members of another group, but the differences in scores on the selection instrument are not manifest in differences in job performance. The guidelines call for investigation of the fairness of selection procedures whenever a selection device has adverse impact.

Table 5.2 EEOC Technical standards Guidelines for the content validation method

	Technical Standard for Content Validation Studies
1.	Consideration must be given to the appropriateness of content validity strategy. Such a strategy is not appropriate when the job tasks represent knowledge, skills, and abilities that an employee is expected to learn on the job. It is also not appropriate for demonstrating the validity of selection procedures that claim to measure traits or constructs such as intelligence, aptitude, personality, common sense, judgment, and leadership.
2.	The job analysis must focus on the important work behaviors, their relative importance across all behaviors, and the products of such work behaviors. To be included in a work sample, the behaviors must be observable, and some aspect of them must be measurable. The work behaviors selected for measurement should be critical and/or important work behaviors that constitute most of the job.
3.	To demonstrate content validity of a selection procedure, it must be shown that the behaviors are a representative sample of behaviors of the job or that the selection procedure offers a representative sample of the work product of the job. For selection procedures measuring a skill or ability, the procedures must closely approximate an observable work behavior or work product. The closer the content and the context of the selection tests are to work samples and work behaviors, the more suitable they are for showing content validity.
4.	Whenever feasible, measurement of the reliability of the selection procedures should be carried out.

Table 5.3 EEOC Technical standards Guidelines for the construct validation method

	Technical Standard for Construct Validity Studies*
1.	The Guidelines recognize that establishment of construct validity is a more complex strategy than either content or criterion-related validity, and that there was, at the time of Guidelines' publication, a lack of literature extending the concept to employment practices.
2.	Therefore, the job analysis must be carried out in a fashion that allows the identification of constructs underlying the important job behaviors. Each construct discovered should be named and defined to distinguish it from all other constructs so discovered.
3.	Selection procedures should then be developed or identified that measure the work behavior constructs. The users must then show that the selection procedures are related to the work behavior constructs and that the work behavior constructs are validly related to the performance of important or critical work behaviors.
4.	The Guidelines allow limited use of construct validity studies. "Until such time as professional literature provides more guidance on the use of construct validity in employment situations, the Federal agencies will accept a claim of construct validity without a criterion-related study... only when the selection procedure has been used elsewhere in a situation in which a criterion-related study has been conducted and the use of a criterion-related validity study in this context meets the standards for transportability of criterion-related validity studies set forth above...."

* see Figure 5.2

Physiological Validation

The validation models identified in the EEOC Guidelines (EEOC, 1978) are based on the American Psychological Association standards for validating educational and psychological tests (A.P.A, 1985; A.P.A., 1987). A major difference when validating physical tests is the use of physiological, not psychological, tasks. Physiological tests differ from educational and psychological tests.

The goal of physiological validation is to match the worker with the physiological demands of the job. An essential element of this process is the quantification of the task's physiological stress. The recent court ruling of Lanning *v.* SEPTA (U.S. 3rd Circuit 1999) gives legal support to physiological validation. The case is discussed in greater detail in Chapter 7 of this State of-the-Art Report (SOAR). A key issue in the Lanning *v.* SEPTA case was setting a valid aerobic fitness cut-score. The recommended cut-score represented a VO_2max of 42.5 ml/kg/min. The court ruled the standard to be unacceptable because the test developers failed to identify the minimum aerobic capacity demanded by the job.

The tradition and "standard practice" used to validate criterion-related physiological tests is to use the metric of the dependent variable (i.e., the criterion test) as the basis for evaluating a subject's work capacity, which is sampled with the predictor test. The metric of the criterion variable has physiological significance. This physiological test validation methodology is clearly illustrated with body composition and VO_2max concurrent test validation research. To illustrate, in 1951, Brozek and Keys (1951) not only reported the concurrent validity coefficient between the predictor test, skinfold fat, and the criterion variable, hydrostatically measured percent body fat, but also published the first regression equation providing a valid model to interpret a subject's skinfold fat measurement by the more meaningful metric of percent body fat. As another example, the maximum treadmill test following a standard protocol is a method of measuring VO_2max. These concurrent validation studies (Bruce, Kusumi, & Hosmer, 1973; Foster, Jackson, & Pollock, 1984; Pollock et al., 1976) published a regression equation with functions to estimate VO_2max (ml/kg/min) from treadmill time. The metric used to interpret aerobic fitness is VO_2max, not elapsed treadmill time. The next section of this chapter examines differences in the validation of physiological and psychological tests.

Differences in Physiological and Psychological Test Validation

Although the psychological-based validation strategies outlined in the EEOC Guidelines are suitable for validating physical tests, there are at least three important differences. These include the test metric used, the work task definition, and the matching of the worker to the demands of the task.

Test Metric—The first major difference between psychological and physiological tests is the test's metric. Typically, the metric of physiological tests is a ratio measurement scale. In contrast, scaling of psychological tests is either ordinal or interval. The units of measurement of physiological tests include percent body fat, oxygen uptake, caloric expenditure, force exerted, pounds lifted, weight load transported, and various types of power output, to name a few. The unit of measurement has physiological significance. In contrast, the unit of measurement of psychological tests is typically an individual's response on a knowledge test or response to some type of scale (e.g., Lickert scale). The unit of measurement on psychological tests is of little importance. This is evidenced by the common practice of transforming scores on psychological tests from the original metric into some form of standard score with a known mean and standard deviation, such as 500 and 100. The person's score is interpreted relative to the mean and standard deviation of the test. In contrast, a physiological test is not only interpreted with the mean and standard deviation of a population, but the value can also have an important physiological meaning. For example, a VO_2max of 20 ml/kg/min not only signifies a person has low fitness by normative standards but also indicates that the person lacks the physiological capacity to perform work tasks with an energy cost that exceeds the person's low aerobic capacity.

Accurate Quantification of Work Demands—A characteristic of physiological test validation is that the physical demands of work tasks can often be objectively measured. This is because of the capacity to define the physical demands of the work task. Extensive physiological research has defined the energy expenditure of a host of occupational, recreational, and fitness tasks by measuring oxygen consumption while doing the tasks (Durnin & Passmore, 1967; Passmore & Durnin, 1955). These energy-cost tables are published in basic exercise physiology texts (Åstrand & Rodahl, 1986; Brooks & Fahey, 1984; McArdle, Katch, & Katch, 1991; Wilmore & Costill, 1994). The forces required to "crack" valves and push or pull objects can be measured with torque wrenches and electronic load cells (Jackson, Osburn, Laughery, & Sekuls, 1998; Jackson, Osburn, Laughery, & Vaubel, 1992). The demands of materials-handling tasks can be defined by weight load, type of lift, lift rate, and distance transported (Jackson, Osburn, Laughery, & Young, 1993a; Waters et al., 1999; Waters, Putz-Anderson, Garg & Fine, 1993). These objective data define the physiological stress demanded by work tasks.

Match the Worker to the Physiological Demands of the Task—A final difference between physiological and psychological test validation is the capacity to match the worker to the physiological demands of the work task. Once the demands of the work task are known, the next step of a physiologically-based validation strategy is to determine if a worker has the capacity to meet the demands of the task. This was the method used to define the minimum energy cost (i.e., VO_2max) required for fire-fighting (Sothmann et al., 1990). This research showed individuals with a VO_2max below 33.5 ml/kg/min were unable meet the demands of firefighting. A goal of ergonomic research has been to define the strength levels needed to do industrial tasks safely (Keyserling et al., 1980; Keyserling, Herrin, & Chaffin, 1980). The next sections of this chapter discuss these methods in more detail.

Physiological Validation—Test Fairness

The goal of a physiological criterion-related strategy is not only to estimate the validity of the test but also determine the minimum physiological level required by the task. A second important element of this approach is the physiological interpretation of the obtained data analyses. Interpretation of the statistical results of validation research with relevant physiological theory and published research provides a scientific rationale to explain the results. Failure to do this leaves the validation results open to question.

An important issue to resolve in a criterion-related study is whether the preemployment test is fair. Unfairness is defined as a situation in which members of a protected group obtain lower scores on a preemployment test than members of another group, but the difference in scores is not reflected in differences in the criterion of job performance (EEOC, 1978). This is called the Cleary test of fairness and is affirmed by showing that the regression line that defines the relationship between the preemployment test and the criterion is common to both groups. The statistical procedure is to test for homogeneity of regression slopes and intercepts (Arvey & Faley, 1988; Jackson, 1989; Pedhauzur, 1997). The literature provides examples of the use of this test (Arnold, Rauschenberger, Soubel, & Guion, 1982; Reilly, Zedeck, & Tenopyr, 1979).

Although the Cleary test may evaluate the fairness of an employment test, the analyses can also provide a physiological interpretation of the employment test. The Cleary test is the method of determining whether a common regression equation can be used to explain the relationship between the predictor and criterion tests of two groups. In physical test validation, the two groups are typically male and female applicants. The data analysis strategy is first to determine whether the two groups share a common regression slope and then decide whether the groups' regression intercepts are within chance variation. Multiple regression is the statistical model used to test for fairness. This multivariate analysis involves dummy-coding the group variable (e.g., female = 0, male = 1) and forming a group by predictor test interaction term (Pedhauzur, 1997). The statistical strategy used is to generate a full multiple regression consisting of the three variables—

1. a predictor test
2. a dummy-coded group variable, and
3. an interaction term, which is the product of the group and test variables.

The next step is to generate two restricted regression models: the first with two independent variables, the group variable and the predictor test; and second, with just the predictor variable. The statistical test used to evaluate group differences in slopes and intercepts is to evaluate changes in R^2 between the full and restricted models. Pedhauzur (1997) outlines these statistical methods and tests of significance. These methods are illustrated next with physiological data. Also shown are the role and importance of the physiological interpretation of the results.

Group Difference in Regression Slopes—A task analysis of freight mover tasks showed that rapidly moving packages from a container to a conveyor belt was a physically demanding task (Jackson et al., 1993a). A work-sample test was developed to duplicate the demands of this repetitive transport task. The task involved moving packages that ranged in weight from about 15 to 80 pounds. The distribution of package weights was representative of the weight distribution encountered by workers. A work-sample test duplicated work demands of the task. Exercise heart rate was measured to ensure the work rate of the simulation test was representative of the actual work rate. The subjects were instructed to work at a brisk rate consistent with their fitness and not to move packages that exceeded their capacity.

Figure 5.3 is the bivariate relationship between the predictor test (sum of isometric strength) and the criterion test (materials transport, expressed in a metric of power output, the pounds of freight transported per minute). The data are contrasted by gender. Analysis of these data showed that male and female regression lines were not parallel. The R^2 change between the full model and restricted model of the strength test and dummy-coded gender variable was 0.04, which was statistically significant ($F_{(1,199)} = 18.96$ $p < 0.01$). The graph shows that the slope for the female subjects (0.534) is more than twice as steep as the slope for male subjetcs (0.208).

A strict interpretation of the Cleary test would indicate that the strength test was unfair, but a physiological interpretation of the data gives a clearer view. Post hoc examination of the data showed that many females could not lift and transport the heavier packages. The lift weight exceeded their strength capacity. The steeper female slope showed that individual differences in strength were more important for females than males. The stronger women could lift the heaviest weight

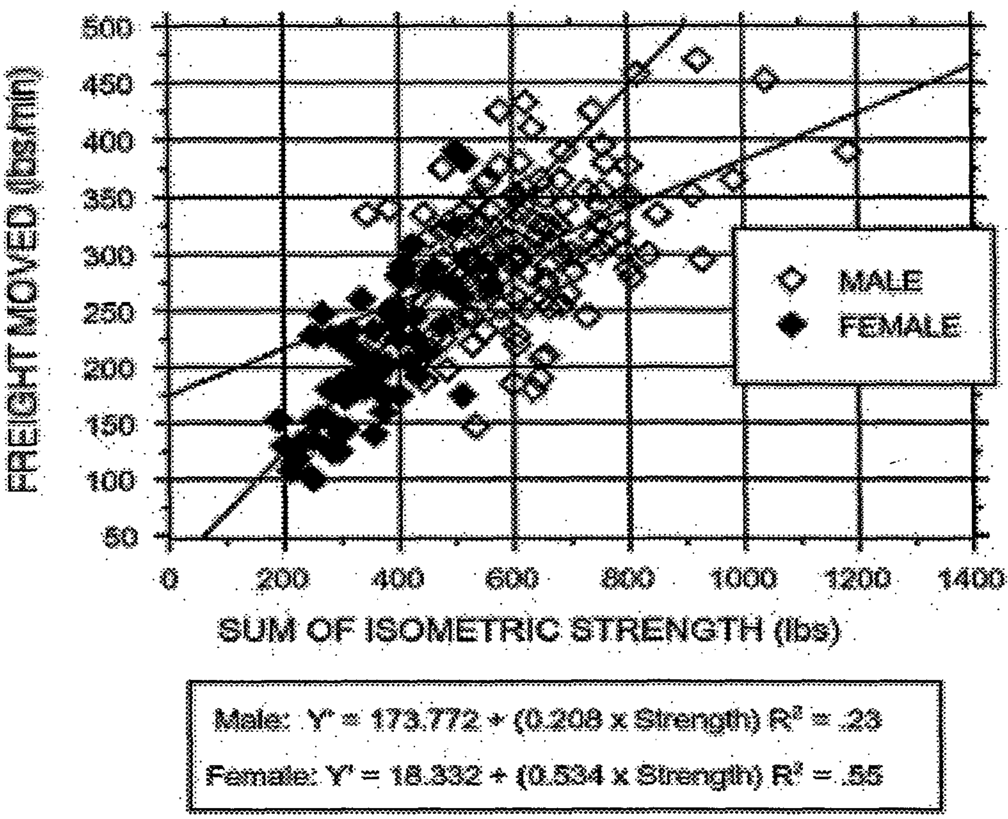

Figure 5.3 Test for fairness, example of significant differences in male and female regression slopes

loads while the weaker women could not. A major determinant of the female capacity to move freight was the subject's strength-dependent capacity to lift heavy loads. In contrast, most men had the physiological capacity to lift and transport the heaviest loads. These physiological data would be important information for setting a cut-score consistent with the demands of the task. The data could also have important ergonomic implications that could lead to job redesign, such as a company policy limiting the weight of packages they would transport.

Intercept Differences—The second part of the Cleary test is to evaluate differences in regression intercepts. Figure 5.4 shows a physiological example of intercept differences in the form of the scatterplot of published male and female body composition data (Jackson & Pollock, 1978; Jackson, Pollock, & Ward, 1980). The independent variable is the sum of seven skinfold measurements, and the dependent variable is percent body fat measured by the underwater weighing method. The figure shows that the slopes of the male and female regression lines are parallel; the differences in slope are within random variation ($F_{(1,675)} = 1.25$; $p > 0.05$). The R^2 difference between the full model and restricted model with gender and the sum of skinfolds was 0.0004. Adding the dummy-coded gender variable to the sum of skinfolds accounted for more than 12 percent of percent fat variance ($F_{(1,675)} = 398.75$; $p < 0.01$). As these data show, the significant intercept difference indicates that for a given score on the predictor test (sum of skinfold fat), the criterion score of one group can be expected to be systematically higher, which in this instance is measured percent body fat. The regression lines differed by an average percent body fat of about 6 percent.

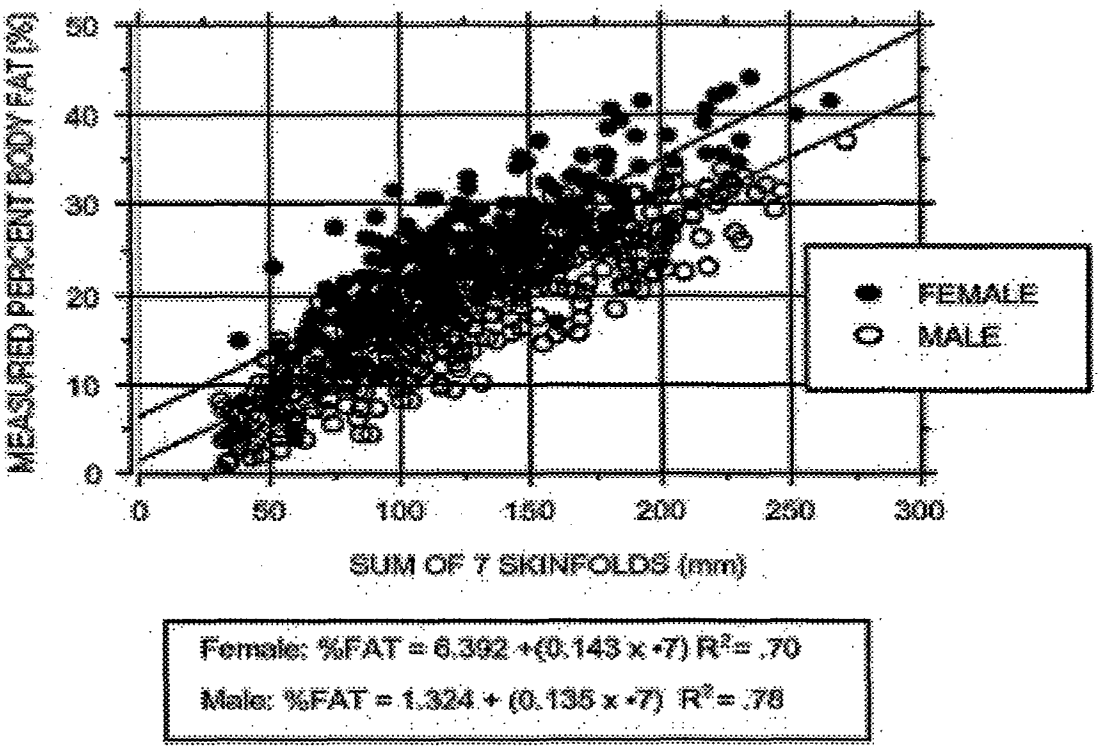

Figure 5.4 Test for fairness, example of parallel regression slopes, but significant differences in male and female regression intercepts

A "blind" application of the Cleary test would indicate that the test was unfair. A physiological interpretation of these results provides a clear rationale for the intercept difference. Skinfold fat measures subcutaneous fat, but the body has two types of fat, subcutaneous and essential fat. Hydrostatically determined percent body fat measures both sources of body fat. It is well established that the essential fat of women is greater by about 7 percent of body mass than that of men (McArdle et al., 1996). The physiological explanation for the gender difference in intercepts can be explained by differences in essential fat.

Although this body composition example does not represent a work-sample test, the use of body composition tests has been an interest of Military researchers (Marriott, 1992). It is well-documented that percent fat is inversely related with strenuous tasks that involve moving the body. This body composition example shows that if percent body fat is used to evaluate male and female performance on common physical tasks (e.g., running, climbing), the test must to be expressed in the physiological metric of percent body fat, not the sum of skinfold fat. In contrast, if the goal is to evaluate fitness rather then the capacity to meet the demands of a work task, gender-based standards are appropriate (Gettman, 1993).

Common Slope and Intercept—The example provided in this section illustrates the homogeneity of male and female regression lines for the predictor and criterion tests. Figure 5.5 gives the scatter plot of the male and female relationship between isometric strength and peak push force. A task analysis showed that push force was a physically demanding task required of workers who moved freight containers (Jackson et al., 1993a). The mean push force of the males was 124.6 (SD = 42.2)

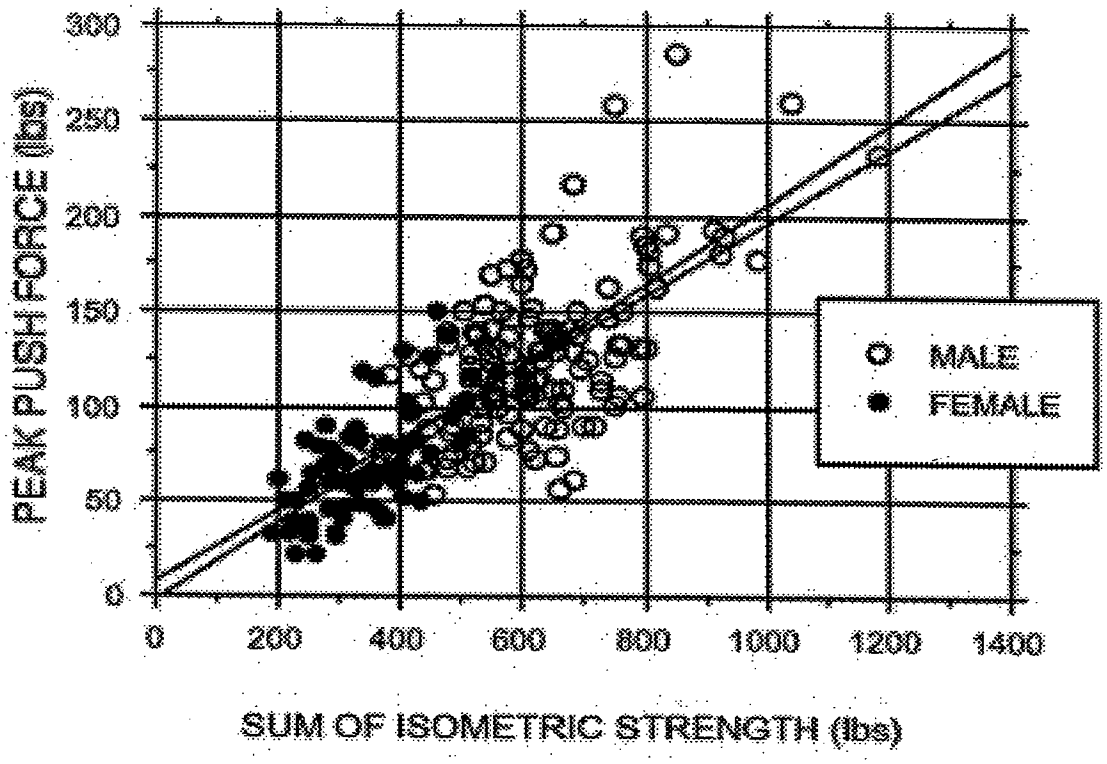

Figure 5.5 Test for fairness, example of homogeneity of male and female regression slopes and intercepts

compared with a mean of 70.0 (SD = 28.3) for the females. This difference was statistically significant ($F_{(1,205)}$ = 99.89; p < 0.01). The figure shows that the male and female regression lines are similar. Statistical analysis showed the slopes ($F_{(1,205)}$ = 1.50; p > 0.05) and intercepts ($F_{(1,203)}$ = 2.00 p > 0.05) of the male and female regression lines were not statistically significant. The group and group-by-strength variables accounted for less then 0.1 percent of the push-force variable. This demonstrated that differences in the regression lines shown in the figure were random variance. This analysis demonstrated that a single regression line can be use to estimate push force from isometric strength, and documented that the gender mean difference in work task performance depended on strength, not gender.

Physiological Validation—Cut-Score

Once the predictor test has been shown to be valid, the next step of a physiological validation strategy is to define performance on the predictor test associated with the desired level of performance on the criterion. An important and often difficult part of this analysis is defining the critical level of performance on the criterion variable. In some instances, a clear definition of an essential task is apparent, for example, lifting a 75-pound industrial valve from the ground to the back of a truck. In other instances, the physiological demands of a task can be difficult to quantify accurately. Shoveling coal is a physically demanding task of coal miners (Jackson & Osburn, 1983), but what level of intensity and duration of shoveling are suitable? Firefighter work simulation tests are timed tests that involve completing several firefighter tasks. Although a firefighter test may be clearly content valid, a more difficult phase of the validation process is to determine the time that signifies successful fire-fighting capacity (Jeanneret & Associates, 1999).

Regression models provide valid statistical methods of estimating physiological capacity, within a defined degree of accuracy, from a predictor test or combination of tests. Simple linear and nonlinear regression models are used with a single predictor test, and multiple regression models are used with several predictor tests (Pedhauzur, 1997). This is a well established physiological test validation method (ACSM, 1991; Åstrand & Ryhming, 1954; Brozek & Keys, 1951; Bruce et al., 1973; Durnin & Wormsley, 1974; Foster et al., 1984; Jackson, 1990; Jackson & Pollock, 1978; Jackson et al., 1980; Pollock et al., 1976). The following provides regression examples of defining physiologically based standards with continuously scaled and pass/fail criterion variables.

Continuously Scaled Criterion—This first example shows the use of simple linear regression to define the strength needed to generate the push force required by a task. The job analysis (Jackson et al., 1993a) showed that one physically demanding job of freight workers was pushing or pulling containers loaded with freight. As part of the job analysis, an electronic load cell defined the peak force required to move freight containers that varied in weight. The subject's peak push force was measured with an isometric push test that simulated the position used to push containers. Figure 5.5 shows the scattergrams with the male and female regression lines. As shown earlier, the difference between the slopes and intercepts of the male and female regression lines were within chance variation which supports the fairness of using a single regression line to define this relationship. The regression equation is—

Push Force Regression Equation $(R = 0.78,\ \ SEE = 29.0\ lbs)$ (1)

$$Push\ \ Force\ \ (lbs) = 2.031 + (0.198 \times Strength)$$

The regression equation provides a valid model for defining the strength needed to generate the push force needed to move containers of the criterion weight. Once this is known, the strength associated with this push force can be determined. To illustrate, assume the criterion push force was defined to be 100 pounds of force. The regression equation shows that a strength score of 495 estimates a push force of 100 pounds.

The goal of a physiological model of validation is to define the minimum physiological capacity demanded by the work task. The regression model provides empirical evidence to define a physiologically defined cut-score within a defined level of probability. Although physiological tests scores typically yield higher criterion-related validity coefficients then psychological tests, they still have substantial prediction errors. Figure 5.6 shows the predictor errors associated with the push force task. Provided is an Altman-Bland plot (Altman & Blaud, 1983; Altman & Blaud, 1986) of the push force data estimated from isometric strength (see Figure 5.5). The Altman-Bland method plots the difference between the residual scores (Y - Y' which is measured estimated push force) by the average of measured and estimated push force. Although the correlation between the criterion, push force, and predictor, isometric strength, was high, 0.78, the Altman-Bland plot shows that defining the physiological criterion is not error free. The variability on the Y axis is defined by the standard error of estimate of the regression analysis, which, in this example, is 29 pounds of push force.

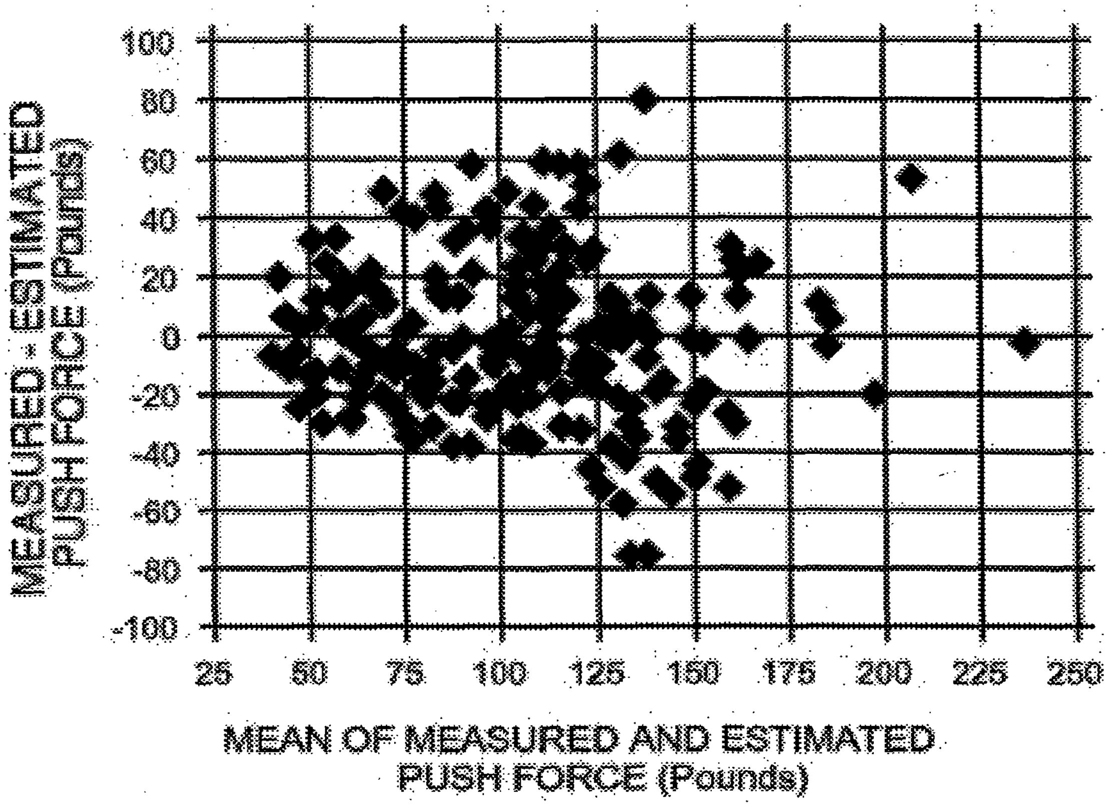

Figure 5.6 Reprinted, by permission, from Altman, D. G., Blandm J. M. (Altman-Bland plot of prediction residuals (measured – estimated) contrasted by the average of measured and estimated maximum push force). pp. 307–310, © by the Lancet Ltd., 1986.

Because the correlation between a predictor variable and the criterion test is always less than 1, there will always be prediction errors. The standard error of estimate provides an estimate of the variation in prediction error. Although it is not possible to define an exact physiologically-based cut-score, it is possible to define a standard with a defined degree of probability. The regression equation (Equation 1) provides a valid model that defines the relationship of strength with push force. As shown earlier, 495 pounds is associated with a push force of 100 pounds. Because the correlation between the two tests is less than perfect and there are prediction errors, only 50 percent of subjects with 495 pounds of strength would be expected to have the capacity to generate 100 pounds of push force. The regression model's standard error of estimate can be used to define the probability that someone, with a given level of strength, would meet the physiologically based standard. Figure 5.7 shows the relationship between level of isometric strength and probability[1] of being able to generate 100 pounds of push force. The probability estimates provide additional data that can be used to define a physiological criterion that is congruent with the criticality of the task, and the mission and unique organizational characteristics.

Pass-Fail Model—Often, the criterion of job performance is scaled as a dichotomous variable. For example, manual lifting tasks are scored pass or fail—the applicant could or could not lift a given weight load (Jackson, Osburn, Loughery & Sekula, 1998; Jackson et al., 1992). Other examples are

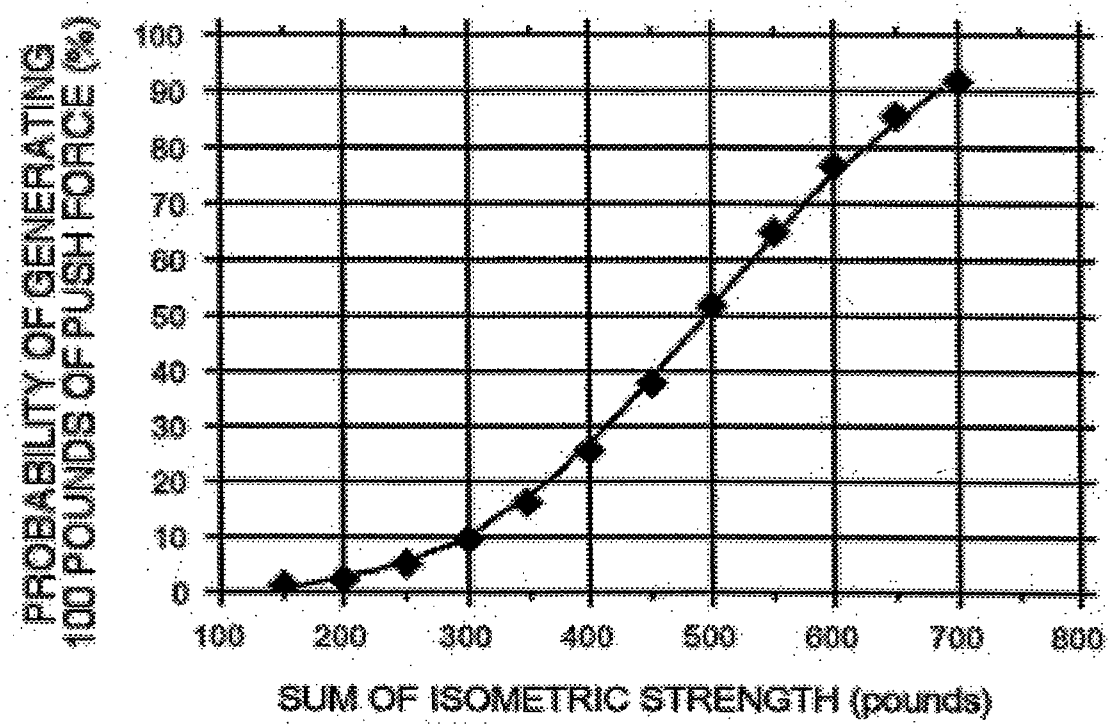

Figure 5.7 Probability of being able to generate 100 pounds for push force for levels of strength

endurance tasks at a constant power output. A manufacturing work task may require a worker to repetitively lift and transport weight loads at a given work rate governed by production speed. Individuals without sufficient physiological capacity would not be able to maintain the set pace. A task documented that refinery workers must close industrial valves during emergencies (Jackson, 1987; Jackson et al., 1992; Osburn, 1977). For some individuals, the task exceeded their physiological capacity and they fatigue quickly. For others, the task was within their physiological capacity. These fit individuals could continue work for extended periods of time. Demanding repetitive tasks at a set power output tend to produce a bimodal distribution—those who have and those who do not have the physiological capacity. This is illustrated in the literature (Jackson et al., 1992).

Logistic regression analysis (Hosmer & Lemeshow, 1989; Pedhauzur, 1997) provides a model to physiologically validate tests when the criterion is a dichotomous variable. Logistic regression, like multiple regression, can use a single independent variable or several independent variables. A logistic regression model estimates the probability of group membership (e.g., criterion variable of pass or fail) given a score or scores on the predictor variable (Pedhauzur, 1997). A public health landmark multiple logistic regression validation study was with the Framingham heart study (Kannel, McGee, & Gordon, 1976). The research objective was to identify and quantify cardiovascular disease risk factors. The logistic analysis not only established that cholesterol, blood pressure, glucose intolerance, and smoking were independent cardiovascular disease (CVD) risk factors, the statistical analysis also produced an equation with a function of estimating the probability of CVD risk for combinations of risk factors. Logistic regression analysis, like regression models with continuous variables, establishes the validity of the independent variable(s) and provides an empirical model for defining the probability of group membership. The application of simple logistic regression analysis is illustrated below with a lifting task.

A task analysis of an oil production plant showed that lifting heavy valves from the floor to knuckle height was an important, physically demanding work task (Jackson, 1998). A work-sample test was developed to simulate the task. The work-sample test involved lifting several loads that varied in weight. The physical dimensions of the lift duplicated the work task. The test was scored pass or fail depending on the subject's ability to complete the lift. The predictor test was the sum of four isometric strength tests, arm, shoulder, torso, and leg strength. The goal of this physiological validation was to define the level of strength required for the lift task.

This validation method is illustrated with three weight loads, 60-, 90-, and 120-pound lifts. These weights represent industrial lifts ranging from moderately heavy to very difficult. The first step in this analysis was to determine whether lift success depended on strength. Table 5.4 provides the means, standard deviations, and sample sizes of the subjects who passed and failed the lift. Analysis of variance showed that lift success depended on strength and documented three, expected trends. First, the number of individuals who could lift the load decreased with the weight load. Next, the Analysis of Variance (ANOVA) documented that lift success for all three weights depended on isometric strength. The means for those who lifted the weight were significantly higher than for those who could not. Third, the mean strength of those who completed the lift increased with the weight load. These trends are consistent with physiological expectations.

Table 5.4 Sample sizes, strength means and standard deviations, and analysis of strength differences of those who could and could not lift the weight

Lift Weight	Lifted Weight		Did Not Lift Weight		ANOVA F-ratio
	N	M ± SD	N	M ± SD	
60-Pound	120	518 ± 197	16	196 ± 66	41.92*
90-Pound	93	579 ± 175	43	233 ± 101	118.89*
120-Pound	71	644 ± 141	65	301 ± 108	250.69*

* P < 0.0001

Figure 5.8 provides a scatter plot of the subjects' strength data contrasted with their 90-pound lift success. This plot shows the group difference in strength documented by the ANOVA but also shows an overlap in the strength of those who passed and failed the lift. Logistic regression analysis provides a model for estimating the probability of success on the criterion variable (i.e., lifting the load) for given levels on the predictor test (i.e., strength) or, in this example, the probability of being able to lift the load for a level of strength. The logistic regression analysis, which agreed with the ANOVAs (Table 5.1), showed that the regression weight for strength was significantly related to the probability of lifting the given weight. The equations for the three lift loads are—

60-pound lift (2)

$$Logit(P) = (0.020 \times Strength) - 3.926$$

90-pound lift (3)

$Logit(P) = (0.017 \times Strength) - 5.689$

120-pound lift (4)

$Logit(P) = (0.023 \times Strength) - 10.334$

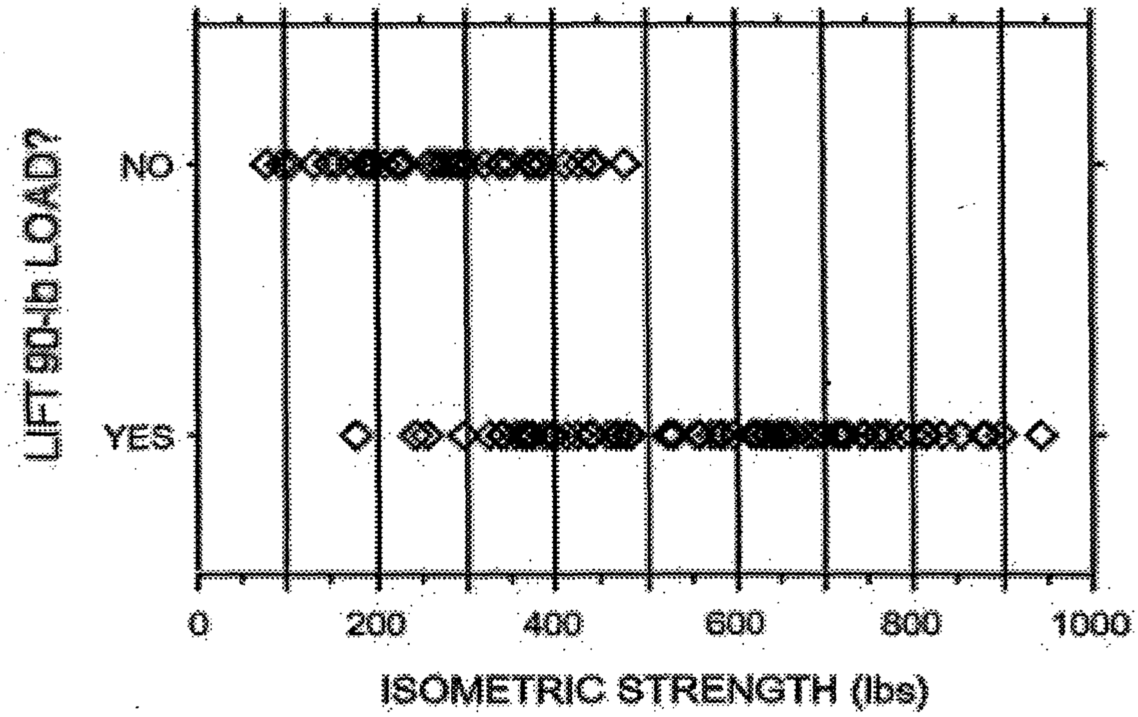

Figure 5.8 Scatterplot of strength test of subjects who could or could not complete a 90-pound lift from floor to knuckle height

Once the logistic equation is defined, Equation 5 estimates the probability of success (Pedhauzur, 1997). The term *e* in Equation 5 is the base of the natural logarithm; a value of Y 2.718. Figure 5.9 graphically shows the probability of success in completing the lift for strength levels.

Logistic Probability Calculation Model (5)

$P = (\frac{e^{a+bX}}{1+e^{a+bX}}) \times 100$

The logistic probability curves clearly show, as would be physiologically expected, that the strength needed to lift the load increases as the lift gets heavier. There is a 50 percent probability, for example, that someone with 200 pounds of strength could lift a 60-pound load. In contrast, only 10 percent of the subjects with 200 pounds of strength would be expected to lift 90 pounds. The likelihood of someone with 200 pounds of strength lifting 120 pounds is 0. The physiological levels needed to be 50 percent confident of lifting the 90- and 120-pound loads are about 350 and 450 pounds of strength, respectively.

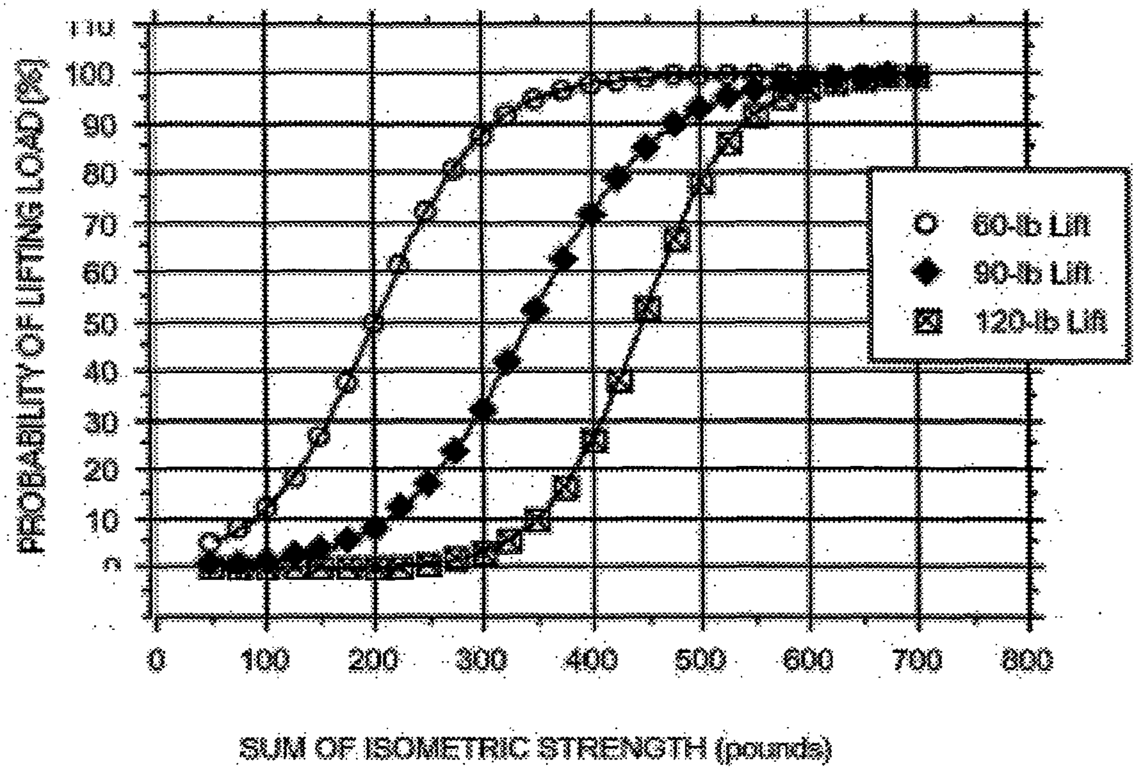

Figure 5.9 Logistic curves of the probability of being able to lift the weight load as a function of lifter strength

Physiological Validation—Matching the Worker to the Job

The goal of physiological test validation is to select workers with the capacity to meet the demands of the job. This is consistent with ergonomic objectives designed to reduce the risk of job-related injuries (Ayoub, 1982). As has been shown in this chapter, the statistical models used to define the physiological stress of the task are less then absolute. This permits latitude in formulating physical cut-scores ranging from lenient to rigorous. The regression statistics, equations, and standard errors provide an empirical base for making the decision.

Although the regression models previously discussed can help define the degree of physiological stress, the difficult task of establishing a suitable cut-score for a criterion remains. The types of job performance criteria listed in the Uniform Guidelines that may be suitable are supervisory ratings, production rate, error rate, tardiness, absenteeism, and success in training. According to the Guidelines, this is not an inclusive list of criteria. Other examples of criteria used to validate physical tests include accidents (Reilly et al., 1979), field performance (Reilly et al., 1979), injury rates (Gilliam & Lund, 2000; Keyserling et al., 1980; Keyserling et al., 1980); lost time due to sickness or injury (Rayson et al., 2000a; Rayson et al., 2000b); and job-related work tasks (Arnold et al., 1982; Jackson, Osburn, & Laughery, 1998; Jackson, Osburn, & Laughery, 1984; Jackson et al., 1992; Jackson, Osburn, & Laughery, 1991; Jackson, Zhang, Laughery, Osburn, & Young, 1993b; Rayson, 2000a; Rayson, 2000b).

A crucial element of any evaluation strategy is the selection rate of a protected group, which, in physical testing, is females. The physiological validation method supplements the process of defining an appropriate cut-score approach with scientific evidence. This validation approach seeks to find the minimum physiological level demanded by the task. The Uniform Guidelines (EEOC, 1978) allow the use of rational judgment in setting a valid cut-score. An objective of the physiological validation process is to provide a scientific explanation of the validation results. Included in this process is the establishment of a sound cut-score. Receiver operator characteristic (ROC) analysis (Hulley, 1988) is one method used to establish physiological cut-scores. It supplements the regression results by defining a cut-score consistent with a strategy of maximizing either test sensitivity or specificity.

A ROC is a graphic analysis used to establish a trade-off between test sensitivity and specificity. If the goal is to maximize test sensitivity, the proportion of true positives (i.e., those who can meet the physiological demands of the work), the ROC would be a plot of test sensitivity by $1 - specificity$, which is the proportion of false positives. False positives are those identified by the test with the physiological capacity to meet the demands of the task but who cannot meet the demands. In this context, the ROC curve provides a rational method of selecting a cut-score based on a balance between high sensitivity and low specificity. The interested reader is directed to another source (Wellens et al., 1996) for the application of ROC analysis for establishing a physiological cut-score. The objective of that study was to find the body mass index (ratio of weight and height) that defined the obesity levels of 25 percent and 33 percent body fat content, determined hydrostatically, for men and women, respectively.

Several factors are considered when establishing physiologically based cut-scores. The following is a nonexhaustive list of conditions that may determine whether a lenient or rigorous cut-score is selected—

- **Adverse Impact**—The first concern is adverse impact. Consideration must be given to the number of the protected group that the standard screens out.
- **Risk of Injury**—Subjecting workers to physical demands increases the risk for work-related injuries. Numerous studies (Cady, Bishoff, O'Connell, Thomas, & Allan, 1979; Gilliam & Lund, 2000; Herrin, 1986; Keyserling et al., 1980; Liles et al., 1984; Snook, Campanelli, & Hart, 1978; Snook & Ciriello, 1991) show that the risk of musculoskeletal injury increases as the demands of the task approach the worker's maximum physiological capacity.
- **Physiological Interpretation of the Validation Results**—An important element of a physical test validation study is to establish the congruence among the validation results, published research, and physiological theory. It is critical to provide a sound physiological explanation of the validation results. Failure to be able to interpret the results by accepted academic standards leaves the decision open to question.
- **Environmental Conditions**—Often, the location at which the validation study is conducted will be different from the work environment. For example, firefighter tests are not administered in burning buildings, the source of demanding work. Environmental conditions (e.g., heat) that increase the demands of the task justify more rigorous standards.
- **Workforce Numbers**—The number of workers available at the work site can affect the rigor of a cut-score. A more lenient standard might be considered when several workers are available to do the work. Although a lenient selection standard would increase the probability that a worker cannot meet the most physical demands of the job (i.e., a false positive), it may

not be a serious problem if others are available to do the work. The stronger workers can help with the most demanding tasks. In contrast, a more rigorous standard might be considered if a worker does not have help.

- **Criticality of the Job**—In some jobs, the failure to meet the demands of a job can be dangerous. The dummy drag test is a common item of a preemployment firefighter test. This is a critical task because the inability to perform it successfully can be life threatening.
- **Workforce Productivity**—Selecting workers with a higher physiological capacity can increase an organization's productivity. The data in Figure 5.3 show that the amount of freight a worker was capable of moving was related to the worker's strength capacity. This was one of the factors considered by a freight company to initiate a preemployment test program.[2]

Published Validation Studies

Although many preemployment tests have been completed, most are not in the published literature. The completed validation study often is a technical report to the governmental agency or private company that funded the project, and many organizations consider these privileged. Hogan (1991b) provides an extensive list of these unpublished reports. The following sections summarize the published validation research.[3]

Outside Craft Jobs

One of the first published concurrent validation studies was for outdoor telephone craft jobs that involved pole-climbing tasks (Bernauer & Bonanno, 1975; Reilly et al., 1979). The issues leading to the development of this study were the large differences between male and female workers in turnover and accident rates. After 6 months, 43 percent of the women left the outdoor craft jobs compared with only 8 percent of the males. More important, women sustained substantially more injuries than men from falls while climbing or working on poles.

An extensive job analysis showed that pole climbing was an essential, physically demanding work task. Bernauer and Bonanno (1975) evaluated the factor composition of 40 tests and anthropometric measures on a sample of 241 job applicants. They developed a six-item battery consisting of reaction time, grip strength, percent body fat, step test performance, balance, and sit-ups. They found that the balance and step tests significantly differentiated successful from unsuccessful students enrolled in pole-climbing school.

Reilly and associates (Reilly et al., 1979) extended this work by completing two concurrent validation studies. In the first experiment, several anthropometric and physical performance tests were administered to 83 male and 45 female candidates for outdoor telephone craft jobs. Two validation criteria were used in this experiment. The first, general task performance, was the average of two supervisor performance ratings of the candidate's performance during the 5-day pole-climbing school. Job analysis data were used to construct the rating scale. The second criterion was a dichotomy of those who were on the job 6 months after placement and those who were not. Using the criterion of general task performance, stepwise multiple regression isolated a three-predictor battery consisting of dynamic arm strength, reaction time, and Harvard bench step time. The analysis yielded a multiple correlation of 0.45. The statistically significant zero-order correlations between the job tenure criterion and these tests were dynamic arm strength, 0.36; reaction time, 0.19; and bench step time, 0.18. Further analysis showed that a common regression line defined male and female performance that met the important criteria of job fairness.

The second experiment used a larger sample of employees who represented the whole company. The criterion of pole-climbing training success was changed to be consistent with changes introduced in the pole-climbing course. The second study included four different criterion measures of job performance—

1. time to complete the pole-climbing school,
2. completion of pole-climbing school (a number withdrew from the course),
3. field observations of pole-climbing proficiency, and
4. accidents for 6 months after entering outdoor craft work.

The second sample consisted of 78 female and 132 male pole-climbing school applicants.

Multiple regression selected a three-item battery consisting of body density estimated from skinfold fat, balance, and an isometric arm strength test. The criterion was time to complete the course.The significant correlations among the three tests and the four criteria were time to complete the course, 0.46; training dropout, 0.38; field observations for the female sample, 0.53; and accidents, 0.15. Further analysis showed that the same regression equation was equally valid for both males and females.

Firefighters

Nearly all major fire departments have a physical ability preemployment test (Landy & Investigator, 1992). Considine and associates (Considine et al., 1976) published the first physical test battery for screening firefighter applicants. The test battery evolved from an occupational task analysis that surveyed, rated, and analyzed 81 tasks performed by firefighters. The authors selected a construct validation strategy. The constructs identified through the task analysis were dynamic strength, static strength, agility, total body coordination, cardiorespiratory endurance, muscular endurance, eye-hand coordination, and total body speed.

The sample of the first study consisted of 191 males who were tested on body composition measures, general physical performance tests, and eight job sample tests. A factor analysis of these data produced three general factors. The factor names and tests representing each factor were factor 1, the ability to handle the body weight measured by percent body fat, obstacle run, and flexed-arm hang; factor 2, muscle power measured by the hose lift, man-lift-and-carry, and stair climb work sample tests; and factor 3, body structure measured by fat-free weight and height.

A major purpose of the second study was to analyze the test battery for racial bias. Based on the results of the first study, nine tests were administered to 165 firefighters and 19 candidates. Data analysis showed that African-American and white subjects did not differ on any of the tests. These data were factor analyzed producing three common factors. The final recommended battery consisted of four work sample tests, and one fitness test; the flexed-arm hang. The work sample tests were modified man-lift-and-carry that simulated rescuing a trapped victim; stair climb that simulated climbing the stairs in a building; obstacle run that simulated moving the body through confined spaces; and hose couple that involved coupling three hoses to a hose couple.

Davis and associates (Davis, Dotson, & SantaMaria, 1982) examined the relationship between simulated firefighting tasks and physical performance measures. The sample consisted of 100 randomly selected men from the population of Washington, DC, firefighters. The physical performance measures included body composition, general fitness, aerobic fitness, and cardiovascular variables. The five work-sample tests came from the job analysis of firefighter work tasks and involved handling a ladder, lifting and transporting a 33.1-kilogram load up five flights of stairs, pulling a 23.5-kilogram hose roll from the ground up to and through the fifth-floor window, carrying and dragging a 53-kilogram dummy down five flights of stairs, and using a sledge hammer to simulate forceful entry.

Canonical correlation showed that two, independent dimensions defined the relationship between the physical performance variables and firefighter work-sample tests. The first canonical dimension (Rc = 0.79) represented a physical work capacity factor that reflected the muscular strength and endurance, and maximal aerobic capacity elements of the simulated work-sample tests. The second dimension (Rc = 0.63) represented a resistance to fatigue factor and the ability to complete the work tasks quickly. Multiple regression selected two physical performance batteries (laboratory and field batteries) to estimate each work-sample dimension. The field test battery for the physical work capacity factor consisted of push-ups, sit-ups, and grip strength. The validity of the field battery (R = 0.73) was lower than the five-item laboratory battery (R = 0.95) that added submaximal oxygen pulse and maximum heart rate to the battery. The three-item field test of the second factor included estimated percent body fat, lean body weight, and VO_2max estimated with a step test (R = 0.77). The laboratory test added maximum heart rate and treadmill performance and increased the validity (R = 0.89) of the resistance to fatigue work sample factor.

The physiological response of fire fighting has been the focus of many investigators. Exercise heart rate responses elicited by simulated and actual firefighting tasks confirmed that these tasks have a significant cardiovascular effect (Barnard & Duncan, 1975; Davis & Convertino, 1975; Lemon & Hermiston, 1977; Manning & Griggs, 1983; O'Connell, Thomas, Caddy, & Karwasky, 1986; Sothmann, Saupe, Jasenor, & Blaney, 1992). In a study during actual fire-suppression emergencies, Sothmann and associates (Sothmann et al., 1992) measured exercise heart rate and oxygen uptake on 10 male fire fighters. Their data showed that firefighters worked at an average of 88 percent (± 6%) of their measured maximum heart rate for an average duration of 15 (±7) minutes. The average energy cost of the firefighter emergency work task was a VO_2 of 25.6 ± 8.7 ml/kg/min, representing an intensity of 63 percent (± 14%) of VO_2max.

Sothmann and associates (Sothmann et al., 1990) examined the relationship between VO_2max and firefighting work tasks. A seven-item, content-valid fire suppression test was administered to 20 experienced fire fighters. The average energy cost of the firefighter simulation tests was 30.5 (± 5.6) ml/kg/min. The work simulation required the firefighters to work at an intensity of 76 percent (± 8) of VO_2max. The correlation between the elapsed time required to complete the firefighter work simulation test and measured VO_2max was -0.55. In a cross-validation study with 32 different male firefighters, successful work simulation performance depended on VO_2max. Of the 32 tested, seven firefighters could not complete the work sample tests. The VO_2max of five of the seven was below 33.5 ml/kg/min.

Highway Patrol Officers

With an increasing number of women seeking employment as highway patrol officers, the objective of the study published by Wilmore and Davis (1979) was to find the minimum physical qualifications and develop a job-related preemployment test. They administered three different batteries of tests to 140 male and 16 female patrol officers. The laboratory and field test batteries included strength, flexibility, body composition, and cardiorespiratory endurance items. The job sample tests included a barrier surmount and arrest simulation, and a dummy drag that simulated dragging an injured victim 50 feet to safety.

The major differences between the field and laboratory batteries were that the 1.5 mile run replaced the maximum treadmill test, and body fat was estimated from skinfolds rather then measured by hydrostatic weighing. The laboratory test battery was significantly correlated with the dummy drag (R= 0.66) and barrier surmount and arrest simulation tests (R= 0.68). Replacing the laboratory tests with the field tests resulted in slightly lower correlations, 0.57 for the dummy drag, and 0.62 for the barrier surmount and arrest simulation tests. Although the fitness tests estimated work simulation test performance, test performance was not related to job performance consisting of supervisor ratings on 16 critical job tasks.

The data analysis showed that the officers were similar to the normal population in strength, body fat, flexibility, and cardiorespiratory endurance. An important result of the study was that the predominantly sedentary nature of the officer's job led to a rapid deterioration in physical fitness following his or her academic training, suggesting the need for an in-service physical conditioning program.

Steel Workers

Arnold and associates (Arnold et al., 1982) developed a preemployment test for selecting entry-level steel workers. The task analysis documented that entry-level steel workers must do several different physically demanding tasks. The investigators used a combination of content-and construct-validation strategies. The job analysis identified the physically demanding work tasks required of the entry-level workers and categorized them by Fleishman's constructs of static strength, dynamic strength, and endurance (Fleishman, 1964). The selected candidate physical performance tests were those that theoretically measured these constructs.

The objective of the study was to determine whether the physical performance tests were related to the work-sample tests developed from the job analysis. The sample included 168 men and 81 women who were in their first 6 months of employment at three different plant locations. The job analysis showed that work tasks differed somewhat across the 3 sites, resulting in 11 work sample tests at 1 site and 12 at the other 2 sites. The average work-sample test performance was the criterion of work performance. In addition to the work-sample tests, each subject completed 10 physical performance tests sampling strength, flexibility, agility, balance, and cardiorespiratory endurance dimensions.

Multiple regression selected the physical performance tests most highly correlated with the work-sample criterion. For all three work sites, arm dynamometer strength was the most important predictor of work-sample test performance. The zero-order correlations between arm strength and work-sample test performance were consistently high—0.82, 0.85, and 0.85 for the three sites. Adding two more tests to the multiple regression models added little to the validity; the multiple correlations for the three predictor models increased to 0.87, 0.88, and 0.89.

The authors completed a utility analysis for the single arm strength test (Hunter, Schmidt, & Hunter, 1979). This analysis involved estimating the money the company would save by hiring workers who could do the work. Utility estimates were based on test validity and the monetary value was related to the variability of work performance. Using 1982 wage standards, Arnold and associates estimated that using the single arm strength test to select employees would lead to a savings of about $5,000 per year for each employee selected. Based on employees hired, the estimated company savings were more than $9 million a year.

Underground Coal Mining

A job analysis showed that the work of underground coal miners was physically demanding and that the work could be represented with four work sample tests (Jackson & Osburn, 1983; Jackson et al., 1991). The first work-sample simulation test, roof bolting, measured maximum isokinetic torque and simulated straightening a steel roof bolt. The block carry test involved lifting, transporting, and placing 82-pound concrete blocks in positions commonly used to build retaining walls in the mine. The shoveling simulation test involved shoveling polyvinyl chloride from the floor over a 3.5-foot wall. Polyvinyl chloride has the same density of coal, and the task was to shovel 800 pounds at a rate consistent with the subject's fitness. The bag carry simulation test measured the number of 50-pound bags that were lifted and transported 9 feet during a 5-minute period.

The four work-sample tests and three isometric strength tests (grip, arm lift, and torso lift) (NIOSH, 1977) were administered to 25 male and 25 female subjects. The validation strategy was similar to that followed by Arnold and associates with steelworkers (Arnold et al., 1982). The correlations between the sum of the isometric strength tests and four work-sample tests ranged from 0.68 for the bag carry test to 0.91 for the roof bolting test. Multiple regression analysis showed that neither gender nor the gender-by-isometric strength interaction accounted for the additional significant variance. This showed that a common male and female regression line defined the relationship between strength and work-sample test performance.

Both exercise heart rate and rating of perceived exertion data showed that the shoveling and bag carry tests had significant aerobic components (Jackson et al., 1991). In addition to the isometric strength tests, the subject's maximal arm cranking oxygen uptake was metabolically determined. The zero-order correlations between the sum of isometric strength and the work-sample shoveling and bag carry tests were higher than the correlations found with arm VO_2max (ml/min). The strength correlations were 0.71 for shoveling and 0.63 for the bag carry test, compared with 0.68 and 0.46 for arm VO_2max (ml/min). Multiple regression analysis showed that arm VO_2max accounted for an additional 9 percent of shoveling variance beyond that of isometric strength but did not account for additional bag carry variance. Polynomial regression analysis showed that the relationship between these two endurance work-sample tests and isometric strength was quadratic, not linear. Strength was more important for differentiating among work sample performance at the lowest levels.

Chemical Plant Workers

Job analyses documented that the physically demanding tasks required of chemical and refining plants workers included cracking, opening, and closing valves (Jackson, Osburn, Laughery, & Vaubel, 1990; Osburn, 1977). Osburn (1977) developed a valve-turning work-simulation test administered on a specially developed ergometer consisting of a disc brake mechanism turned by a 12-inch value handwheel. The unit was calibrated to a power output of 1,413.5 foot-pounds/minute. The objective of the work-sample test was to complete 250 revolutions in 15 minutes. The job analysis showed this level of work would open or close 75 percent of the emergency valves in 15 minutes.

The distribution of the valve-turning test was bimodal. Physically fit workers easily completed the 15-minute test, but the test was too demanding for many, who stopped before reaching 50 revolutions (Jackson et al., 1990). The test elicited maximal cardiovascular responses in many applicants (Osburn, 1977). This result led to a second study designed to determine whether isometric strength tests validly predicted valve-turning performance (Jackson, 1987; Jackson et al., 1992). The valve-turning work-sample test, and three isometric strength tests (grip, arm lift, and torso lift) were administered to 26 men and 25 women. The zero-order correlation between the tests was 0.82. Because of the bimodal shape of the valve-turning distribution, a logistic regression model (Pedhauzur, 1997) defined the probability of completing the test by levels of isometric strength. The logistic equations and probability curves are published (Jackson et al., 1992).

In a second study, a task analysis questionnaire completed by operators at a major chemical plant identified valve cracking as the most physically demanding work task (Jackson et al., 1990). An electronic load cell measured the peak cracking torque on 217 randomly selected valves in the plant. The sampled valves included those with horizontal and vertical orientations, positioned close to the ground and overhead, those in awkward or hard to reach positions, and valves of various sizes. The results of this biomechanical job analysis showed that 100 pounds of force applied to the end of a 36-inch valve wrench generated sufficient torque to crack 93 percent of the plant valves.

A valve-cracking work-sample test simulated cracking valves in eight different ways. The eight cracking torques were obtained by varying the action (push and pull), direction (horizontal and vertical), and height (high and low). A computerized torque wrench measured the torque applied to four nuts placed in vertical and horizontal positions at two heights.

The valve-cracking test and isometric strength tests (grip, arm lift, and torso lift) were administered to 118 men and 66 women. The intercorrelations among the eight measures of valve-cracking torque were high, ranging from 0.66 to 0.89. Because of the high intercorrelations, the eight valve-cracking scores were averaged and used as the work-sample measure. The correlation between the sum of the three isometric strength tests and average valve-cracking torque was 0.65. A logistic regression equation (Pedhauzur, 1997) defined a probability model for estimating the chances of generating the 100-pound criterion for levels of isometric strength. These data are published elsewhere (Jackson et al., 1992).

Electrical Transmission Lineworkers

Doolittle and associates (Doolittle et al., 1988) developed a preemployment test for selecting electrical transmission lineworkers. The study included an extensive job analysis of electrical transmission lineworker jobs. The initial stage of the task analysis surveyed workers using scales designed to answer three questions—

1. How often was each task performed?
2. How much time was spent completing each task?
3. How physically demanding was each task for the individual?

The identified critical, physically demanding tasks were studied in detail to define the forces needed to perform them safely and efficiently. This involved defining standard anatomical movements for lifting, pushing, and hoisting; measuring the masses lifted and forces exerted; and estimating the metabolic costs of various work tasks.

Using the task analysis data, 5 strength tests that duplicated the muscular actions were selected and administered to 48 incumbents. The tests required the subject to move a weight that represented loads that linemen moved. The weights ranged from 7 to 61 kilograms. The final two tests selected were chin-ups and VO_2max estimated from bench stepping and exercise heart rate. The seven tests were combined into a single performance measure. Criterion-related validity was examined by comparing physical test performance with two criteria, supervisor ratings and accident rates. The crew chiefs confidentially evaluated each incumbent on the following six dimensions of job performance—

1. productivity,
2. working with others,
3. supervision,
4. safety,
5. physical ability, and
6. technical skills.

The correlations between the composite physical test criteria of supervisor ratings and lost work days because of on-the-job injuries averaged over 5 years were 0.59 and 0.46.

Diver Training

Two validation studies (Gunderson, Rahe, & Arthur, 1972; Hogan, 1985) were designed to estimate successful completion of Military underwater diver training programs. Gunderson and associates (Gunderson et al., 1972) used successful completion of underwater demolition training as the criterion of performance. They found a multiple correlation of 0.54 between success defined by the completion of training and five variables, squat-jumps, pull-ups, sit-ups, body weight, and the Cornell Medical Index. Using these tests, they predicted about 70 percent of those who passed training.

Hogan (Hogan, 1985) used 46 male, naval personnel who volunteered for diver training. The first criteria was success included nine performance rating scales that reflected physical condition, swimming training, leadership potential, teamwork, and overall performance. The second criteria was successful completion of training. The predictor measures included 3 anthropometric measurements and 23 fitness tests. Hogan reported a multiple correlation of 0.63 between the average performance rating and three physical tests, 1-mile run, sit and reach, and muscular endurance measured with an arm ergometer. The multiple correlation between these three tests and successful completion of the course was 0.64. Hogan suggested that the validity coefficients were likely an overestimate because of an unfavorable ratio of the number variables and subjects (Pedhauzur, 1997).

Demanding Military Jobs

The U.S. Military Services examined methods of matching enlisted personnel with physically demanding jobs. The U.S. Air Force adopted a pre-induction dynamic one-repetition maximum (1–RM) strength test (Ayoub et al., 1982). The U.S. Army and U.S. Navy examined the relationship between body composition variables and physically demanding work tasks (Marriott & Grumstrup-Scott, 1992).

The U.S. Air Force developed a Strength Aptitude Test (SAT) to match the general strength abilities of individuals with the specific strength requirements of U.S. Air Force jobs filled by enlisted personnel (Ayoub et al., 1982). The U.S. Air Force SAT measures the subject's voluntary 1–RM lift to a height of 6 feet. The SAT starts with a 40-pound lift. The lift load is increased by 10 pounds until the subject reaches his or her maximum voluntary lift or a maximum weight of 200 pounds. The SAT is administered to U.S. Air Force recruits as part of their pre-induction physical examination. Each enlisted U.S. Air Force career field has a prerequisite SAT cut-score.

An area of concern expressed by the Committee on Military Nutrition Research of the Institute of Medicine, National Academy of Sciences, is the role body composition plays in physical performance. This relationship is important not only for making decisions about acceptance or rejection of recruits for the Military Service but also for retention and advancement while in the Service (Marriott & Grumpstrup-Scott, 1992). Hodgdon and associates (Hodgdon, 1992) examined the relationship between body composition, fitness, and materials-handling tasks required of naval enlisted men. The two materials-handling tasks were the maximum box weight that could be lifted to elbow height and the total distance a 34-kilogram box could be carried during two, 5-minute workouts. The variables most highly correlated with maximum box lift were push-ups (r = 0.63) and fat-free mass (r = 0.80). The variables most highly correlated with the box carry test were push-ups (r = 0.56), 1.5-mile run time (r = -0.67), and fat-free mass (r = 0.44). Fat-free mass was highly correlated with muscular strength measures, suggesting the possibility of using fat-free mass as an approximation of general strength in job assignment.

Vogel and Friedl (Vogel & Friedl, 1992) examined the relationship between body composition and absolute lifting capacity. They reported significant correlations between maximum lifting capacity and fat-free mass for male and female soldiers. Although they did not test for homogeneity of male and female regression lines, they published separate equations for men and women.

A limitation of Military testing programs is the lack of job-related materials-handling performance tests. While recognizing the need to develop content-valid tests, the Committee on Military Nutrition Research concluded that there was a direct relationship between Military materials-handling tasks and fat-free mass. In view of this relationship and the lack of job-related tests, the Military should seriously consider establishing a minimum standard for fat-free mass (Marriott & Grumpstrup-Scott, 1992). Such a recommendation might be implemented for the Military, but using body composition variables in pre-employment tests in the private sector would likely meet an immediate legal challenge.

Rayson and associates (Rayson et al., 2000a; Rayson et al., 2000b) completed a major criterion-related validation study for the British army. They examined the effectiveness of the British army's Physical Standards for Recruits (PSS(R)) in predicting criteria measuring recruit success in basic training. The PSS(R) consisted of tests measuring body mass, body composition, strength, and endurance. The criteria included—

1. four representative Military tasks (RMT) consisting of a single lift, carry, repetitive lift, and loaded march,
2. the days lost to injury and sickness during basic training,
3. degree of success of basic training, and
4. job performance ratings by self, peer, and supervisor.

The PSS(R) tests were administered to more than 1,000 recruits (770 males and 239 females) prior to starting basic training, and the army job performance criteria were obtained at the end of basic training.

The PSS(R) tests correctly predicted outcomes on the RMTs for 74.9 percent of the recruits, of which 58.7 percent were true positives and 16.2 percent were true negatives. Of the 25.1 percent misclassified, 15.5 percent were false positives and 9.6 percent were false negatives. The false negatives were those recruits predicted by the PSS(R) tests to fail the four RMTs when they did pass the tasks. Although data were not presented, the authors indicated that most of the female misclassifications were false positives, "...women being incorrectly accepted rather than incorrectly rejected from the army." A significant relationship was found between training outcome and passing the PSS(R) tests. Additionally, the PSS(R) tests were significantly related to days lost because of injury and sickness during basic training. Those recruits who failed their selection outcome lost a median of 2 days compared with no days for the recruits who passed. Although not statistically significant, the performance ratings of those who failed the selection tests were consistently lower then those who passed the tests. The authors concluded that the PSS(R) were valid, useful predictors of British army performance..

Manual Lifting Tasks

Manual lifting tasks are common elements of many jobs. Manual lifting tasks have been studied extensively. The reason for this popularity is the large number of job that include materials-handling tasks and the injury risk associated with lifting. It is estimated that about 50 percent of all industrial back injuries are caused by lifting, and about 67 percent of the injuries are caused by lifting loads that are too difficult for industrial workers (Snook et al., 1978).

An established ergonomic injury-reduction strategy is to match the worker with the demands of the lifting task. One major approach is to engineer the stress out of the task. This approach defines the lift weights that are within the physiological capacity of most industrial workers (Ayoub, 1982). The first research-based strategy used psychophysical methods to define the lift weight perceived as acceptable to 75 percent of industrial workers. Snook and associates (Snook & Ciriello, 1974; Snook & Ciriello, 1991; Snook, Irvine, & Bass, 1970) published separate standards for males and females. The maximum acceptable lift weight for females was about 50 percent of the lift weights for males. A newer strategy is the use of the NIOSH multiplicative equations (NIOSH, 1981; Waters et al., 1993) that consider several different lift difficulty parameters. The NIOSH equations extend the Snook and associates' psychophysical methodology by also using biomechanical and physiological criteria to define recommended weight of lift (RWL). The newest NIOSH equation (Waters et al., 1993) defines a RWL that would be acceptable to 75 percent of the female industrial population. Using the 75th percentile female as the RWL criterion produces a conservative estimate. The RWL for the common floor to knuckle lift at a frequency of one lift every 30 minutes, for example, is only 10 kilograms or 22 pounds (Waters et al., 1993).

The NIOSH equation focuses on job design, i.e., defining a RWL for most male (99 percent) and female (75 percent) industrial workers for all ages in the workforce. A limitation of the NIOSH equation is that it does not consider individual differences in physiological capacity of workers. Many common materials-handling tasks exceed the NIOSH equation's RWL estimates. The second ergonomic method of matching the worker with the demands of job is to select individuals with the physiological capacity to do the job with a margin of safety (Ayoub, 1982; Keyserling & al., 1980; NIOSH, 1977).

The content-validation method is often used to validate materials-handling tests. A content-valid test would be to have the applicant perform the task, e.g., lift a 90-pound jackhammer and transport it a specified distance. Although this type of test would be content valid, it has two limitations. First, it is not possible to determine one's maximum capacity. Second, motivated applicants without the physiological capacity demanded by the task place themselves at risk of injury (Ayoub, 1982). One of the first ergonomic approaches used to overcome these limitations was to use isometric strength tests that duplicated the position assumed by the worker to do the lift. These position-specific strength data were used to determine whether an applicant had sufficient strength capacity to do the work with a margin of safety (Keyserling & al., 1980; Keyserling et al., 1980).

Gilliam and Lund (2000) examined the effects on work-related injuries of physiologically matching workers to the demands of the job. Isokinetic strength was measured on 365 applicants for truck driver and dockworker jobs. The isokinetic data were used to generate a Department of Labor Dictionary of Occupational Titles strength rating. This rating was used to select applicants who matched the physical demands of the job. Of the 365 applicants, 276 matched the job demands and were hired. The 89 applicants who did not match were not hired. Those hired were significantly stronger then those who were not hired. In addition, those not hired were significantly heavier then those hired. Those not hired were 44 pounds heavier then the new hires.The injury rates of the strength-matched new hires were compared with historical data on workers matched for employment duration. The overexertion injury rates to the knees, shoulders, and back were 1.04 for the strength-matched workers compared with 16.7 for the non-matched workers, suggesting that preemployment screening is effective in reducing injury. Although not examined, these results also suggest that body composition may also have been a factor. A strength-weight profile of weaker and heavier versus stronger and lighter suggest a difference in percent body fat. The stronger-lighter profile is consistent with a lower percent body fat, which also might have been an injury risk factor.

Another physiological approach to matching the worker to the demands of the job is to use standard strength tests to assess an individual's physiological capacity and use regression models to define the probability of being able to complete a lift (Jackson & Sekula, 1999; Jackson, Borg, Zhang, Laughery, & Chen, 1997). This approach was used to study hospital workers involved with lifting and transporting patients. An analysis of hospital jobs documented that patient lifting was a demanding lift task (Jackson, Osburn, Laughery, Young, & Zhang, 1994). Patient lift tasks are a major source of injury to the lifter (Garg & Owen, 1992). The lift dimensions of the most common single-person patient lift were used to devise a work-sample lift test. The most common patient lift task is lifting a patient who is sitting in a chair.The simulated lift test consisted of lifting a box from a height of 53 cm to a height of 48 cm. The hand position at the start of the lift was at a height that the lifter would grab a patient sitting in a chair. The lift task consisted of lifting seven loads ranging in weight from 15 to 90 pounds. The subjects lifted those loads that were within their capacity and rated lift difficulty with Borg's CR–10 psychophysical scale (Borg, 1982; Borg, 1998). Logistic regression analysis of the data on 58 female and 33 male subjects showed that the capacity to complete a lift depended on the lifter's physiological capacity sampled by his or her isometric strength and fat-free mass. Further analyses showed that the subject's CR–10 rating of each lift was significantly correlated with isometric arm, shoulder, torso, and leg strength, and fat-free weight.

The results of the patient lift study suggested that lift weight and the physiological capacity of the lifter could be used to develop a generalized lift model. The second study examined the role of lift load, strength, and gender on psychophysical lift capacity (Jackson, 1999). A floor-to-knuckle lift test was administered to 209 men and 181 women. The task involved lifting loads ranging from 22 to 143 pounds. The subject started with a light lift load and continued to lift heavier loads until either the heaviest load was lifted or the subject failed the lift. The load increased at a linear rate of 11 pounds. After each completed lift, the subject rated the lift difficulty with Borg's CR–10 scale (Borg, 1998). The subject's physiological strength capacity was measured with basic isometric strength tests (Baumgartner & Jackson, 1999). Each subject's dynamic lift profile was defined with a power function regression equation using the completed lift weight as the independent variable and the CR–10 rating as the dependent variable. Using the power function regression equation, one lift weight and the associated CR–10 rating were randomly selected for each subject. This created a distribution of lift weights and associated psychophysical ratings ranging from very easy to the maximum within the subject's psychophysical capacity. Multiple regression provided an equation with a function to estimate psychophysical lift difficulty from lift load, strength, and the gender-by-weight load interaction. The multiple correlation for the model was 0.81, with a standard error of 1.7 CR–10 units. The derived equation provided a model that defined the psychophysical lift demands of common industrial weight loads for individuals who differed in physiological capacity.

The psychophysical modeling of industrial lift tasks not only provides evidence concerning an individual's probability of being able to complete a lift but also psychophysical stress. The psychophysical demand of a lift task is related to the risk of back injury (Herrin, 1986; Liles, 1984;

Snook et al., 1978). Lifting loads psychophysically judged to be difficult increases the risk of injury. Psychophysical ratings provide an index of relative demand for the individual. Resnik (1995) presents preliminary data showing that Borg's psychophysical rating can be interpreted by the physiological significant scale of percentage of maximum capacity. With a sample of 254 male and 354 female subjects, a correlation of 0.91 was obtained between Borg's CR–10 rating and the subject's maximum function lift capacity (Sekula, Jackson, & Laughlin, under review). Maximum functional lift capacity was the subject's percentage of maximum lift, where maximum lift represented the weight load equal to the subject's Borg psychophysical CR–10 rating of 10. A regression equation was developed to convert CR–10 ratings into the metric of percentage of max. The standard error of estimate for the linear equation was 8.5 percent max. This research could provide researchers with the capacity to interpret psychophysically defined lift loads with the well-established physiological intensity metric of percentage of maximum capacity.

Summary

In summary, the Uniform Guidelines require validity studies to be carried out whenever there is a need to continue selection practices that lead to adverse impacts. Three types of validity studies are recognized: content-validity, criterion-related validity, and construct-validity studies. The guidelines require all validity studies to be carried out in a responsible, scientifically sound manner, and call for the use of good judgment in the implementation of selection procedures. The EEOC is waiting for developments in the field before it completely endorses construct-validity studies. A major difference in physical test validation is the use of physiological rather then psychological tests.The goal of physiological validation is to define the physiological capacity needed by a worker to perform the work demanded by the task. Principal features of the physiological validation approach are the use of a physiological metric to quantify test performance and the interpretation of validity results using relevant physiological research and theory. These data are used to develop physiologically sound cut-scores. Although numerous physical test validation studies have been completed, most are not published. The results of those published shows that physical tests can be used to select workers with the physiological capacity to do demanding jobs. Ergonomic research shows that selecting workers with the physiological capacity to do the work reduces the risk of work-related injuries.

Endnotes

1. The probability can be estimated with the following equation: $z = \frac{Y-criterion}{standard\ error\ of\ estimate}$, where Y' is the estimated criterion score and the criterion is the desired value, in this example, 100. Once the z-score is obtained, a table of normal curves can be used to estimate the proportion of subjects that can be expected to exceed the criterion for a given strength level.
2. Personal communication between A. Jackson, University of Houston, and Dr. John Hater of the Fedex Corporation. Engineers used the power output data in Figure 5.3 to estimate expected changes in productivity produced by changes the physiological capacity of the workforce.

3. This review was initially published in 1994 by one of the authors of this chapter (Jackson, 1994) and expanded to include studies published since that time.

References

1. A.P.A. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.
2. A.P.A. (1987). *Principles for the validation and use of personnel selection procedures.* Washington, DC: Division of Industrial-Organizational Psychology, American Psychological Association.
3. ACSM. (1991). *Guidelines for exercise testing and prescription.* (3rd ed.). (Vol. 4). Philadelphia, PA: Lea and Febiger.
4. Altman, D. G., Bland, J. M. (1983). Measurement in medicine: The analysis of method comparison studies. *The Statistician, 32,* 307–317.
5. Altman, D. G., Bland, J. M. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet, I,* 307–310.
6. 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.
7. Arvey, R. D., & Faley, R. H. (1988). *Fairness in Selecting Employees.* (2nd ed.). Reading, MA: Addison-Wesley Publishing Co.
8. Åstrand, P.-O., & Rodahl, K. (1986). *Textbook of work physiology.* (3rd ed.). New York: McGraw-Hill.
9. Åstrand, P.-O., & Ryhming, I. (1954). A nomogram for calculation of aerobic capacity (physical fitness) from pulse rate during submaximal work. *Journal of Applied Physiology, 7,* 218–221.
10. Ayoub, M. A. (1982). Control of manual lifting hazards: II. Job redesign. *Journal of Occupational Medicine, 24,* 688–676.
11. Ayoub, M. M., Denardo, J. D., Smith, J. L., Bethea, N. J., Lambert, B. K., Alley, L. R., & Duran, B. S. (1982). *Establishing physical criteria for assigning personnel to Air Force jobs* (Contract No. F49620–79–C0006). Lubbock, TX: Texas Tech University.
12. Barnard, R., & Duncan, H. W. (1975). Heart rate and ECG responses of firefighters. *Journal of Occupational Medicine, 17,* 247–250.
13. Baumgartner, T. A., & Jackson, A. S. (1999). *Measurement for evaluation in physical education and exercise science.* (6th ed.). Dubuque: Wm. C. Brown.
14. Bernauer, E. M., & Bonanno, J. (1975). Development of physical profiles for specific jobs. *Journal of Occupational Medicine, 17,* 22–33.
15. Borg, G. (1982). A category scale with ratio properties for intermodal and interindividual comparisons. In H. G. Geissler & P. Petzold (Eds.), *Psychophysical judgment and the process of perception.* Berlin: VEB Deutscher Verlag der Wissenschaften.
16. Borg, G. (1998). *Borg's perceived exertion and pain scaling method.* Champaign: Human Kinetics.
17. Brooks, G., & Fahey, T. (1984). *Exercise physiology: Human bioenergetics and its applications.* New York: John Wiley and Sons.

18. Brozek, J., & Keys, A. (1951). The evaluation of leanness-fatness in man: Norms and intercorrelations. *British Journal of Nutrition, 5*, 194–206.

19. Bruce, R. A., Kusumi, F., & Hosmer, D. (1973). Maximal oxygen intake and nomographic assessment of functional aerobic impairment in cardiovascular disease. *American Heart Journal, 85*, 546–562.

20. Cady, L. D., Bishoff, D. P., O'Connell, E. R., Thomas, P. C., & Allan, J. H. (1979). Back injuries in firefighters. *Journal of Occupational Medicine, 21*, 269–272.

21. Considine, W., Misner, J. E., Boileau, R. A., Pounian, C., Cole, J., & Abbatieilo, A. (1976). Developing a physical performance test battery for screening Chicago fire fighting applicants. *Public Personnel Management, 5*, 7–14.

22. Davis, J. A., & Convertino, V. A. (1975). A comparison of heart rate methods for predicting endurance training intensity., *Medicine and Science in Sports, 7*, 295–298.

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

24. Doolittle, T. L., Spurlin, O., Kaiyala, K., & Sovern, D. (1988). Physical demands of lineworkers. *Proceedings of the Human Factors Society, 32nd Annual Meeting, 32*, 632–636.

25. Durnin, J. V. G. A., & Passmore, R. (1967). *Energy, work and leisure.* London: Heinemann Educational Books, LTD.

26. Durnin, J. V. G. A., & Wormsley, J. (1974). Body fat assessed from total body density and its estimation from skinfold thickness: Measurements on 481 men and women aged from 16 to 72 years. *British Journal of Nutrition, 32*, 77–92.

27. EEOC. (1978). Uniform Guidelines on employment selection procedures. *Federal Register, 43* (38289–28309).

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

29. Foster, C., Jackson, A. S., & Pollock, M. L. (1984). Generalized equations for predicting functional capacity from treadmill performance. *American Heart Journal, 107*, 1229–1234.

30. Garg, A., & Owen, B. (1992). Reducing back stress to nursing personnel: An ergonomic intervention in a nursing home. *Ergonomics, 35*, (11), 1353–1375.

31. Gettman, L. R. (1993). Chapter 19 Fitness Testing. In *Resource manual for guidelines for exercise testing and prescription* (2nd ed., pp. 229–246). Philadelphia: Lea & Febiger.

32. Gilliam, T. B., & Lund, S. J. (2000). Injury reduction in truck driver/dock workers through physical capability new hire screening. *Medicine and Science in Sports and Exercise, 32*, (5) (Supplement), S126.

33. Gunderson, E. K. E., Rahe, R. H., & Arthur, R. J. (1972). Prediction of performance in stressful underwater demolition training. *Journal of Applied Psychology, 56*, pp. 430–432.

34. Herrin, G. D., Jaraiedi, M., Anderson, C. K. (1986). Prediction of overexertion injuries using biomechanical and psychophysical models. *American Industrial Hygiene Association Journal, 47*, 322–330.

35. Hodgdon, J. A. (1992). Body composition in the military services: standards and methods. In B. M. Marriott & J. Grumstrup-Scott (Eds.), *Body composition and physical performance: Applications for the military services* (pp. 57–70). Washington, DC: National Academy Press.

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

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

38. Hogan, J. C. (1991b). Chapter 11 Physical Abilities. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of Industrial and Organizational Psychology* (2nd ed., Vol. 2, pp. 743–831). Palo Alto, CA: Consulting Psychologist Press, Inc.

39. Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic tegression.* New York: John Wiley & Sons.

40. Hulley, S. B., Cummings, S. R. (1988). *Designing clinical research.* Baltimore, MD: Williams & Wilkins.

41. Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin, 86*, 721–735.

42. Jackson, A. S. (1987). *Validity of isometric strength tests for predicting work performance of refinery workers.* Houston: Shell Oil Company.

43. Jackson, A. S. (1989). Chapter 9: Application of regression analysis to exercise science. In M.J. Safrit & T.M. Wood (Eds.), *Measurement concepts in physical education and exercise science.* Champaign: Human Kinetics.

44. Jackson, A. S., Blair, S. N., Mahar, M. T. , Wier, L. T., Ross, R. M. , Stuteville, J. E. (1990). Prediction of functional aerobic capacity without exercise testing. *Medicine and Science in Sports and Exercise, 22*, 863–870.

45. Jackson, A. S. (1994). Chapter 3 Preemployment Physical Evaluation. *Exercise and Sport Science Reviews, 22*, 53–90.

46. Jackson, A. S., Osburn, H. G., Laughery, K. R., Sekula, B. K. (1998). *Revalidation of methods for preemployment assessment of physical abilities at Shell Western Exploration and Production, Inc., and CalResources LLC.* Houston, TX: University of Houston.

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

48. Jackson, A. S., Borg, G., Zhang, J. J., Laughery, K. R., & Chen, J. (1997). Role of physical work capacity and load weight on psychophysical lift ratings. *International Journal of Industrial Ergonomics, 20*, 181–190.

49. Jackson, A. S., & Osburn, H. (1983). *Preemployment physical test development for coal mining technicians.* Technical Report to Shell Oil Co. Houston: Shell Oil Company.

50. Jackson, A. S., Osburn, H. G., & Laughery, K. R. (1984). Validity of isometric strength tests for predicting performance in physically demanding jobs. *Proceedings of the Human Factors Society 28th Annual Meeting, 28*, 452–454.

51. Jackson, A. S., Osburn, H. G., Laughery, K. R., & Vaubel, K. P. (1990). *Validation of physical strength tests for the Texas City Plant—Union Carbide Corporation.* Houston: Center for Psychological Services.

52. Jackson, A. S., Osburn, H. G., Laughery, K. R., & Vaubel, K. P. (1992). Validity of isometric strength tests for predicting the capacity to crack, open and close industrial valves. *Proceedings of the Human Factors Society 36th Annual Meeting, 1*, 688–691.

53. Jackson, A. S., Osburn, H. G., Laughery, K. R., & Young, S. L. (1993a). *Validation of physical strength tests for the Federal Express Corporation.* Houston: Center of Applied Psychological Services, Rice University.

54. Jackson, A. S., Osburn, H. G., Laughery, K. R., Young, S. L., & Zhang, J. J. (1994). *Patient lifting tasks at Methodist Hospital.* Houston: Center of Applied Psychological Services, Rice University.

55. Jackson, A. S., Osburn, H. G., & Laughery, S., K.R. (1991). Validity of isometric strength tests for predicting endurance work tasks of coal miners. *Proceedings of the Human Factors Society 35th Annual Meeting, 1*, 763–767.

56. Jackson, A. S., & Pollock, M. L. (1978). Generalized equations for predicting body density of men. *British Journal of Nutrition, 40*, 497–504.

57. Jackson, A. S., Pollock, M. L., & Ward, A. (1980). Generalized equations for predicting body density of women. *Medicine and Science in Sports and Exercise, 12*, 175–182.

58. Jackson, A. S., Zhang, J. J., Laughery, K. R., Osburn, H. G., & Young, S. L. (1993b). *Final report: Patient lifting tasks at Methodist Hospital.* Houston: Center of Applied Psychological Services, Rice University.

59. Jeanneret & Associates, I. (1999). *Evaluation of physical ability and written tests for entry-level firefighters*. St. Paul, MN: City of St. Paul Fire Department.

60. Kannel, W. B., McGee, D., & Gordon, T. (1976). A general cardiovascular risk profile: The Framingham study. *American Journal of Cardiology, 38*, 46–51.

61. Keyserling, W. M., et al. (1980). Establishing an industrial strength testing program. *American Industrial Hygiene Association Journal, 41*, 730–736.

62. Keyserling, W. M., Herrin, G. D., & Chaffin, D. B. (1980). Isometric strength testing as a means of controlling medical incidents on strenuous jobs. *Journal of Occupational Medicine, 22*, 332–336.

63. Landy, F. J., & Investigator, P. (1992). *Alternatives to chronological age in determining standards of suitability for public safety jobs: Volume I: Technical report.* The Pennsylvania State University: Center for Applied Behavioral Sciences.

64. Lemon, P., & Hermiston, R. T. (1977). The human energy cost of firefighting. *Journal of Occupational Medicine, 19*, 558–562.

65. Liles, D. H., Deivanayagam, S., Ayoub, M. M., Mahajan, P. (1984). A job severity index for the evaluation and control of lifting injury. *Human Factors, 26*, 683–693.

66. Manning, J., & Griggs, T. (1983). Heart rates in fire fighters using light and heavy breathing equipment: Similar near-maximal exertion in response to multiple work load conditions. *Journal of Occupational Medicine, 25*, 215–218.

67. Marriott, B. M., Grumstrup-Scott, J. (Eds.) (1992). *Body composition and physical performance: Application for the military services*. Washington, DC: National Academy Press.

68. McArdle, W. D., Katch, F. I., & Katch, V. L. (1991). *Exercise physiology: Energy, nutrition, and human performance*. (3rd ed.). Philadelphia: Lea & Febiger.

69. McArdle, W. D., Katch, F. I., & Katch, V. L. (1996). *Exercise physiology: Energy, nutrition, and human performance*. (4th ed.). Philadelphia: Lea & Febiger.

70. Meyers, D. C., Gebhardt, D. L., Crump, C. E., & Fleishman, E. A. (1984). *Factor analysis of strength, cardiovascular endurance, flexibility, and body composition measures* (Tech. Rep. R83–9). Bethesda, MD: Advanced Research Resources Organization.

71. NIOSH. (1977). *Preemployment strength testing*. Washington, DC: U.S. Department of Health and Human Services.

72. NIOSH. (1981). *Work practices guide for manual lifting*. Washington, DC: U.S. Department of Health and Human Services.

73. O'Connell, E., Thomas, P., Caddy, L., & Karwasky, R. (1986). Energy costs of simulated stair climbing as a job-related task in fire fighting. *Journal of Occupational Medicine, 28*, 282–284.

74. Osburn, H. G. (1977). *An investigation of applicant physical qualifications in relation to operator tasks at the Deer Park Manufacturing Complex.* Houston, TX: Employee Relations, Shell Oil Company.

75. Passmore, R., & Durnin, J. V. G. A. (1955). Human energy expenditure. *Physiological Review, 35,* 801–840.

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

77. Pollock, M. L., Bohannon, R. L., Cooper, K. H., Ayres, J. J., Ward, A., White, S. R., & Linnerud, A. C. (1976). A comparative analysis of four protocols for maximal treadmill stress testing. *American Heart Journal, 92,* 39–42.

78. Rayson, M., Pynn, H., Rothwell, A., Nevill, A. (2000a). Validation of job-related fitness tests. *Medicine and Science in Sports and Exercise, 32,* (5) (Supplement), S126.

79. Rayson, M. P., Pynn, H., Rothwell, A., Nevill, A. (Ed.). (2000b). *The development of physical selection procedures for the British Army. Phase 3: Validation.* London: Taylor and Francis.

80. Reilly, R. R., Zedeck, S., & Tenopyr, M. L. (1979). Validity and fairness of physical ability tests for predicting craft jobs. *Journal of Applied Psychology, 64,* 267–274.

81. Resnik, M. L. (1995). The generalizability of psychophysical ratings in predicting the perception of lift difficulty. *Proceedings of the Human Factors Society,* 679–682.

82. Rummel, R. J. (1970). *Applied factor analysis.* Evanston: Northwestern University Press.

83. Sekula, B., Jackson, A. S., Laughlin, M. S. (Under Review). *Measuring percentage of maximal functional lift capacity: Calibration of Borg's CR–10 scale.*

84. Snook, S. H., Campanelli, R. A., & Hart, J. W. (1978). A study of three preventive approaches to low back injury. *Journal of Occupational Medicine, 20,* 478–481.

85. Snook, S. H., & Ciriello, B. M. (1974). Maximum weights and work loads acceptable to female workers. *Journal of Occupational Medicine, 16,* 527–534.

86. Snook, S. H., & Ciriello, V. M. (1991). The design of manual handling tasks: revised tables of maximum acceptable weights and forces. *Ergonomics, 34,* 1197–1213.

87. Snook, S. H., Irvine, C. H., & Bass, S. F. (1970). Maximum weights and work loads acceptable to male industrial workers. *American Industrial Hygiene Association Journal, 31,* 579–586.

88. Sothmann, M. S., Saupe, K., Jasenor, D., & Blaney, J. (1992). Heart rate response of firefighters to actual emergencies. *Journal of Occupational Medicine, 34,* 797–800.

89. Sothmann, M. S., Saupe, K. W., Jasenof, D., Blaney, J., Donahue-Fuhrman, S., & Woulfe, T. (1990). Advancing age and the cardiorespiratory stress of fire suppression: Determining a minimum standard for aerobic fitness. *Human Performance, 3,* 217–236.

90. Vogel, J. A., & Friedl, K. E. (1992). Army Data: Body composition and physical capacity. In B. M. Marriott & J. Grumstrup-Scott (Eds.), *Body composition and physical performance: Applications for the military services* (pp. 89–104). Washington, DC: National Academy Press.

91. Waters, T. R., Baron, S. L., Piacitelli, L. A., Anderson, V. P., Skov, T., Haring-Sweeney, M., Wall, D. K., Fine, L. J. (1999). Evaluation of the revised NIOSH lifting equation. *Spine, 24,* 386–395.

92. Waters, T. R., Putz-Anderson, V., Garg, A., & Fine, L. J. (1993). Revised NIOSH equation for the design and evaluation of manual lifting tasks. *Ergonomics, 7,* 749–766.

93. Wellens, R. I., Roche, A. F., Khamis, H. J., Jackson, A. S., Pollock, M. L., & Siervogel, R. M. (1996). Relationships between body mass index and body composition. *Obesity Research, 4,* 35–44.

94. Wilmore, J. H., & Costill, D. L. (1994). *Physiology of sport and exercise.* Champaign, IL: Human Kinetics.

95. Wilmore, J. H., & Davis, J. A. (1979). Validation of a physical abilities field test for the selection of state traffic officers. *Journal of Occupational Medicine, 21*, 33–40.

REPORT DOCUMENTATION PAGE

1. Report Date (DD MM YY)
21 Sep 2000

2. Report Type
New

3. DATES COVERED (from - to)
1996 to Feb 2000

4. TITLE AND SUBTITLE
Physical Test Validation for Job Selection

5a. Contract Number:
5b. Grant Number:
5c. Program Element: 63706N
5d. Project Number: M0096
5e. Task Number: 002
5f. Work Unit Number: 6716
5g. IRB Protocol Number:

6. AUTHORS
James Hodgdon & Andrew S. Jackson

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Health Research Center
P.O. Box 85122
San Diego, CA 92186-5122

8. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES)
Chief, Bureau of Medicine and Surgery
M2
2300 E Street, NW
Washington, DC 20372-5300

9. PERFORMING ORGANIZATION REPORT NUMBER
Report No. 01-12

10. Sponsor/Monitor's Acronyms(s)
BUMED

11. Sponsor/Monitor's Report Number(s)

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

13. SUPPLEMENTARY NOTES
Chapter in Constable & Palmer (eds), The Process of Physical Fitness Standards Development, 2000, 139-177. Human Systems IAC SOAR

14. ABSTRACT (maximum 200 words)

This State of the Art Review (SOAR) chapter examines the issues related to physical test validation for job selection. The chapter is divided into three major sections. The first examines issues and accepted methods of test validation. The focus is on the interpretation of the Equal Employment Opportunity Commission guidelines as they relate to test validation. The sanctioned validation methods are content validity, criterion-related validity; and construct validity. The measurement theory used to evaluate the quality of employment tests is based on the American Psychological Association standards for validating educational and psychological tests. A major difference in physical test validation is the use of physiological rather then psychological tests. The second section of the chapter examines the differences between physiological and psychological test validation. The goal of physiological validation is to define the physiological capacity needed by a worker to perform the work demanded by the task. Principle features of physiological validation approach are the use of a physiological metric to quantify test performance and the interpretation of validity results with relevant physiological research and theory. The final section of the chapter is a review of published employment validation research of physical tests.

15. SUBJECT TERMS Physical fitness Tests, Employment Selection, EEOC Guidelines, Test Validation, Content Validity, Criterion-related validity, construct validity

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b.ABSTRACT	b. THIS PAGE			Commanding Officer
UNCL	UNCL	UNCL	UNCL	39	19b. TELEPHONE NUMBER (INCLUDING AREA CODE) COMM/DSN: (619) 553-8429