Psychological Testing in Personnel Selection, Part II: The Refinement of Methods and Standards in Employee Selection

By Wesley A. Scroggins, PhD, Steven L. Thomas, PhD, and Jerry A. Morris, PsyD

This paper is the second of a three part series that examines the development of selection testing. Part I focused on the historical development of personnel selection testing from the late 19th century to the present, with particular attention given to personality testing. Attention was given to the efforts of early industrial psychologists that shaped and defined the role of testing in the scientific selection of employees. Part II examines the development of methods and standards in employment testing with particular emphasis on selection validity and utility. Issues of selection fairness and discrimination in selection are explored as they relate to psychological testing. Part III explores the development and application of personality testing. The transient nature of models of personality is noted and current paradigms and the utility and fairness of personality testing for modern organizations are discussed.

During the first hundred years of scientific management and personnel selection, psychological testing became a powerful and institutionalized tool that was broadly applied. Universities and trade schools, managers, engineers, psychologists, and government officials recognized the importance of using scientific methods and tools to manage human capital, the economics of business, and the national defense. Industrial and organizational psychologists emerged as the preeminent players in the development and dissemination of these tools and the evaluation of the tools’ real and perceived efficiency. This article reviews the methods, measures, and standards that have emerged to evaluate these tools.

The Issue of Validity

Efforts to develop and create tests to assist managers with personnel selection have ranged from the absurd to the hopeful and into the stage of continued refinement and
growing efficiency. The scientific principles underlying the selection tests have also evolved from simple and crude to complex and powerful. Psychologists have known for some time that ineffective tools and recurring fads are not uncommon within the field of psychological testing, particularly when it comes to the manner in which tests are used. Consequently, psychologists have often cautioned consumers to be wary of the variable quality of what is purported to be psychological science.

In the early part of the 20th century, psychology usurped the use of scientific methods to improve performance, selection, motivation, and strategic training from engineering management and had eclipsed scientific management in terms of contributing to the national defense through selection test development. The assessment tools were designed with the hope of facilitating optimal personnel development and utilization of human capital. Using scientific principles and quantitative methods to conserve and develop optimal human resources by forecasting behavioral tendencies over extended periods proved to be a considerable challenge. The challenge was met with variable success, as the opportunities for error were great. Psychologists had captured the testing agenda, but they would struggle mightily to identify instruments that were accepted and widely judged to be useful.

The most important issue in HR selection testing is determining a test’s validity. The actual definition of validity can vary depending on the circumstances, the specific tools used, and the application. For most selection purposes, however, a selection test is valid if the characteristic(s) it is measuring is related to the requirements and/or some important aspect of the job the test taker is being evaluated to perform. Test scores only have meaning if the test is valid, and a test is valid if there is a link between the test score and job performance. The degree to which an employment selection test has validity tells the testing entity what it can conclude or predict about someone’s job performance from his or her test scores. A test’s validity is established for a specific purpose, and it may not be valid for purposes other than those that it has been validated to measure.

Important issues of validity that are discussed in most HR management textbooks are described in Section 1607.5 of the U.S. Equal Employment Opportunity Commission’s (EEOC) 1978 Uniform Guidelines on Employee Selection Procedures as criterion-related validity, content validity, and construct validity. Criterion-related validity is the correlation or other statistical relationship between selection test score (the predictor) and job performance (the criterion). If those who score low on a test also perform poorly (and visa versa), the test is said to have high criterion-related validity. Content-related validation is a demonstration that the content of the test reflects important job-related behaviors and measures important job-related knowledge or skills. Construct-related validity is evidence that a test measures the constructs or abstract characteristics that are important to successful performance of the job. For psychological tests used in selection, a test’s criterion-related validity is usually the variable of interest to researchers, and it is the validity coefficient—the actual correlation coefficient between a test score and some job performance criterion—that is referred to when validity is discussed in HR literature.
Having evidence of the validity of selection tests is essential for any organization using such tools. Collecting these data is the principal way companies demonstrate that they have met the Uniform Guidelines' requirements should hiring procedures result in adverse impact (i.e., disproportionate hiring outcomes) against protected groups.

Organizations can obtain evidence of criterion-related validity using two basic strategies. One strategy is to conduct a concurrent validation study in which predictor test scores are collected from current job incumbents and the results are correlated with either current measures of job performance or file records of past performance. A second strategy is a predictive validation study in which predictor test scores are obtained from job applicants at the time of hire. After the new employees have had sufficient time to learn the job, job performance measures are obtained, and then the correlation between the predictor test and job performance scores is determined. The problem with an organization conducting its own validation study is that it can be time-consuming (especially if a predictive design is used), it can be costly, and it requires large sample sizes to yield reasonable levels of reliability. Because of these problems, only fairly large organizations and those with fairly frequent hiring have the resources and sample sizes necessary for conducting validity studies.

A third method for establishing selection test validity is to rely on validity generalization, which is often used when organizations want to use a test for selecting individuals for a specific job and choose to use a test purchased from a commercial vendor. The EEOC's Uniform Guidelines and related standards require that there be evidence of transportability (or generalizability) when purchased tests are used. Transportability refers to the necessity that validation procedures used must be consistent with accepted standards, that the test is fair, and that there is job similarity or that a job analysis has indicated that the validity studies for the job relative to the related instrument are substantially similar in terms of ability and work behavior.

Validity evidence can include a demonstration that the specific job for which a test is used to select personnel is very similar to the job for which the test was originally developed. Determining the degree of similarity requires that an organization conduct systematic job analyses as way of identifying the tasks, duties, responsibilities, and work conditions of a job and the knowledge, skills, and abilities required to succeed in the performance of the job.

Additional evidence of the validity of an employee selection test is the validity coefficient, which is the correlation coefficient between an employee's selection test score and one or more relevant measure of job performance. The greater the validity coefficient, the more valid the test is and the greater the usefulness of the test in the selection process. Acceptable validities for a single test are usually in the range of .21 to .35; for multiple tests, somewhat higher ranges (above .35) are customary for considering tests to be valid.

The most important property of a personnel selection test is predictive validity, or the ability to predict job performance, learning, and success. The predictive validity coefficient is a proxy for a full-blown predictive validity study and gives a fair approximation of the degree to which the benefits of using a given test to hire more
effective employees exceed the costs of developing, purchasing, and/or administering a newer test.11

In the past two decades, various selection tools have been shown to increase an organization's ability to hire more productive employees and realize increases in output, the monetary value of output, and the learning of job skills.12 On the other hand, Meehl has described the horrors of using humanistic, rather than scientific, approaches to analyzing and predicting human behavior in industrial and vocational settings during the 1950s.13 The cross-situational validity and generalization of ability and performance predictive validities explored in this era remain cyclically elusive and are greatly discussed and reexamined to this day. This gave way to a period when industrial and vocational psychologists thought that predictive validities were situational specific and that test validities varied from city to city and area to area.14 The theory of situational specificity remained dominant in personnel psychology until the late 1970s, when it was discovered that most of the differences across studies were due to statistical and measurement artifacts related to simple sampling error variation caused by the use of small samples.15

Early on, the field of personnel selection testing struggled with the effects of sample sizes of fewer than 100 on validity correlations.16 Because validity coefficients produced from small samples may be sample-specific, it is possible that much of the association between a predictor and a job performance criterion can be reduced to statistical variation due to sample fluctuations.17 Eventually, powerful meta-analytical techniques were used to combine validity estimates across studies and correct the effects of sampling error. The results convincingly showed that situational specificity or the idea that the validity of a given procedure varied in different settings for basically the same job was untrue.18 Therefore, due to the problems associated with small-sample validation studies and the time and cost associated with large in-house studies, many organizations resorted to using commercially validated tests that were often validated for significantly different uses and populations.

As psychologists turned to the development and evaluation of tests used in business and industry, several problematic issues emerged. Problems included complaints that the particular items of a test lacked relevance (i.e., poor face validity) to important job-related attributes. Test items, regardless of predictive validity, will not be perceived to be fair and will not be accepted by applicants if the items are perceived to be unrelated to the knowledge, skills, and abilities important to the job. Research has consistently shown that applicants favor tests that they perceive to be job-related.19 Applicants have been shown to react more negatively to tests they perceive to be unrelated to the job when they fail to achieve a score that would get them hired20 and that applicants often attribute their failure to the instrument or organization rather than to themselves.21

A concept related to face validity is procedural justice. When selection tests have poor face validity and applicant acceptability, the likelihood that applicants will have negative perceptions of procedural justice within the organization's selection system increases. In other words, applicants are more likely to feel that the organization's selection system is unfair. Face validity clearly has an impact on applicants' perceptions.
of the organization, procedural justice, fundamental fairness, and self-perception after they take a selection test (i.e., personal impact).

**Group Differences in Testing**

A second problematic issue in psychological testing has been that different demographic groups of people tend to perform differently when taking the same tests. Explanations for why this is so are always controversial, but systematic group differences on criterion measures of performance clearly exist for many psychological tests used for prediction.\(^{22}\) Many experts argue that psychology's struggle with ethics and cultural sensitivity has clashed with intellectual honesty in this controversial area.\(^{23}\) Kimble has described the reticence to come to grips with this issue as anti-intellectualism masquerading as human sensitivity.\(^{24}\) Group differences on ability and achievement tests are not artifacts, and it is now clear that certain groups, such as Asian Americans, typically score higher than the general population taken as a whole on standardized aptitude and achievement measures, while other minority groups, such as Hispanics and African Americans, score lower.\(^{25}\)

Since by some estimates only 15-20% of all companies use cognitive ability testing for selection purposes, it seems fairly clear that group differences in test performance is a major factor in many organizations' decision not to use cognitive ability tests.\(^{26}\) In 1971, the U.S. Supreme Court ruled in *Griggs v. Duke Power* (401 U.S. 424) that when a selection test produce adverse impact against protected group members, the company using that test must be able to defend the practice by showing that using the test is a "business necessity" for the operation of the business. This and similar decisions by other courts have clearly had a damping effect on the willingness of companies to engage in cognitive ability testing because increasingly narrow interpretations of what constitutes a business necessity sufficient to justify disproportionate outcomes now require companies to show that no other acceptable selection alternative exist.\(^{27}\)

The fact that some groups perform systematically better or worse than other groups on some psychological tests is particularly troublesome because mental ability tests are known to be good predictors of job performance, especially with regard to the acquisition of job knowledge.\(^{28}\) Since the potential contributions of mental ability testing are constrained, some discussion of validity issues related to the complex and interesting variance of expert opinion in this area is necessary.

Many experts and personnel selection specialists believe that test validity can be attenuated or even sacrificed to reduce adverse impact. Often, a practitioner is faced with a choice among tests having very different costs, degrees of validity, and fairness.\(^{29}\) The *Uniform Guidelines* provide guidance on making such choices: When two procedures are available that are valid and reliable and that serve the company's interest in efficient and trustworthy workmanship, the company should use the procedure that has been demonstrated to have the "lesser adverse impact."\(^{30}\)

The selection professional is faced with the obligation of developing a test that has high validity and minimal or no adverse impact in an environment where racial, gender, and social groups may have different ability distributions and in which poor
selection outcomes would adversely impact the prospects of the hiring organization. Consequently, the procedure known as the *Golden Rule*, which consists of selecting items on the basis of reduced adverse impact and compromising construct validity (and reliability), has become widely held.

Some have described modern personnel selection test choice or development as involving the method of weighing the degree of adverse impact against the strength of evidence of validity. If a test has only minor adverse impact, a demonstration requiring only minimal evidence of validity is commonly considered sufficient. Others have turned to employment interview methods that generally have less validity than do general ability tests, or to clinical approaches to selection decisions as ways of avoiding adverse impact. Some experts have pinned accounts of being exhorted by a court to select an alternative test with less adverse impact even when lower validity would be the result. Other experts do not subscribe to the theory that a valid test has adverse impact, or that validity has to suffer to avoid adverse impact.

**Decision Making and Test Use**

Companies have used a number of strategies to deal with disproportionate outcomes from selection testing. The simplest way to find the most productive workforce and to meet the legal requirements of the Civil Rights Acts of 1964 and 1991 is to employ a pass-fail cutoff that divides a pool of applicants into accepted and rejected groups and then to apply a secondary screening method to ensure that the company does not underutilize protected group members. This is a simple strategy that only requires the company to manipulate a cutoff score so that sufficient numbers of protected group applicants fall above the accepted minimum score to fill available slots. However, this technique requires using secondary tools of sometimes low validity to sort accepted applicants and can result in less than optimal selections. In any event, based on the 1982 decision in *Connecticut v. Teal* (457 U. S. 440), this method does not absolve organizations of the responsibility for validating the initial test even if the bottom-line impact favors protected groups.

A company using a valid test might meet legal requirements by taking the top scorers on a test scores until they reach a sufficient number of minorities to comply with the *80 percent rule*. It has been shown that the economic value of hiring an employee scoring at one standard deviation above the mean on some predictive measures can be as much as 40% higher than hiring an individual who scored at the mean of the predictor.

Though this top-down method of employee selection is much more efficient than using a cutoff score for optimizing human capital and hiring individuals who will be able to acquire job knowledge and be productive while complying with the law, it may be impractical. In the past, when organizations address affirmative action issues, they may incur huge opportunity costs that can have dramatic impacts on their financial well-being. Employers have sought to minimize these opportunity costs by being sensitive and socially conscious about community standards and fairness in employment and addressing social ills while using quotas to select the top candidates.
from very distinct groups. However, deciding in advance how many nonminorities to select before shifting to selecting protected group members to meet the 80 percent rule under a quota system is explicitly prohibited by the Civil Rights Act of 1991. Therefore, just as many valid tests with proven economic impact have been taken out of the hands of organizations, so too has this tool.

So what is an employer who wants to be a good citizen and sees the value of equal rights for all, in order to comply with legal requirements relative to employee selection and meet its fiduciary responsibility to maximize shareholders' wealth? One solution that emerged was race-norming. This was a process in which test scores were adjusted on the basis of ethnic background by converting raw scores into percentiles based on the distribution of scores within each group to limit the effects of adverse impact and maintain larger portions of minorities in applicant pools. For instance, a nonminority applicant with a mean score for nonminorities on an IQ test may score 100, while a minority individual with a mean score for minorities may score 90 on that same IQ test. Both scores would be adjusted to a percentile score of 50, and the nonminority and minority candidates would have an equal chance of being selected. This solution was actually promoted by the U.S. government in the 1970s and was a fairly typical way of dealing with fairness in hiring problems.

The debate over race-norming—perhaps the best example of this conflict between intellectual honesty and cultural sensitivity—came to a head in 1989 a blue-ribbon committee of the National Academy of Sciences concluded that the U.S. Department of Labor's practice of race-norming employment tests was "scientifically justified" because "modest validities of the General Aptitude Test Battery cause selection errors that weigh more heavily on minority workers than on majority workers." Oddly enough, a reversal that came just two years later when Congress added a ban on race-norming and race-conscious score adjustments to the Civil Rights Act of 1991 drew little opposition from the scientific community that had originally subordinated the utility of test validity to ethnical sensitivity, apparently out of a concern for fairness toward nonminority candidates.

Under race-norming the difference in group performance among races is recognized as a scientific fact, but the value of fairness and equal opportunity for all races is also viewed as important. Large employers seeking efficiency in personnel selection are generally in the position of needing tests to identify quality additions to their workforce and to increase the productive capability of the organization. However, employers also face the reality that tests, because of real differences in performance between racial and even gender groups, will produce adverse impact in the current legal sense. Companies must then be concerned with underutilization and compliance with the requirements of the Uniform Guideline's 80 percent rule and hiring additional applicants from protected groups as well as their ability to defend the validity and job-relatedness of their selection procedures.

Another approach that has emerged to deal with the problem of differential group performance on valid tests is banding. Banding is based on a thread of psychometric theory that holds that small differences in test scores are quite possibly due to measurement error and, therefore, these differences within an acceptable band
are considered irrelevant for interpretation and decision-making purposes. Proponents of this theory favor developing bands that are determined by the psychometric properties of the test—notably the test's level of reliability—within which all those scoring in the range are considered qualified for hire and each person within a band is treated as if he or she has the same score and likelihood for successfully performing the job.

Banding is a controversial theoretical approach. Some critics have pointed out that banding threatens to commit the “fallacy of composition” while attempting to reconcile the need for minimizing adverse impact at the potential sacrifice of some test validity. Even proponents of banding have acknowledged that there is a loss of predictive validity that is best accessed with top-down selection methods. Others recognize that even the best selection instruments have some degree of unreliability and that small differences in test scores are often not statistically significant. Those inclined to accept banding for personnel selection testing analysis believe it can significantly reduce adverse impact with only minor effects on utility. However, the success of banding is very much dependent on test reliability. Using banding with highly reliable tests that generate adverse impact will result in a considerable loss of utility with little reduction in adverse impact.

Perspectives on Selection Fairness and Testing

There are many different perspectives from which to view fairness in personnel selection, and each has implications for psychological testing and decision making. Dreher and Sackett identified five models of fairness, including pure quota and culture-free models in which fairness in both is defined based on how well selection ratios represent protected group proportions in a relevant labor market. Other definitions, such as the models developed by Cleary and by Thorndike, incorporate the issues of single-group validity versus differential validity and differential prediction. In these models, the fairness of a selection test is determined by examining the regression line describing job performance as a function of test scores, which is the line of best fit through a scatterplot of data points that show an individual's test scores on the horizontal axis and job performance measures on the vertical axis. Typically, a test will only be considered to be fair when the regression lines for minority and nonminority groups are the same. When regression lines are not the same for minority and nonminority groups, then using a test in the same way for both would not be considered fair.

Some have argued that a valid test has no adverse impact if it is within the subset of fair tests. In classic models of test fairness, when tests are compared and all tests in the cadre are proven fair under established fairness criterion, the most valid test will be the one with the least adverse impact. Schmidt and Hunter have concluded that when a personnel selection test passes the test of fairness and the majority group scores more favorably, there is an inverse relationship between the validity of the test and its adverse impact.
Proponents of this theory hold that current validity coefficients are computed based on the validity of the measures aggregating across both protected and nonprotected populations. This is said to ignore group membership, so critics argue that a more meaningful measure of validity would be within-group validity. With regard to Cleary’s model of fairness, if the regression line demonstrating performance as a function of the predictor measure for protected group members is identical to the line for nonprotected group members, the test is fair. If the regression lines for protected and nonprotected groups are not identical, then the regression line is biased and each person in both groups will receive an unbiased prediction of performance only if separate regression lines based on group membership are used. If a test is fair, group membership is irrelevant. But if a test is biased, separate regression lines should be used—which is equivalent to using race as a predictor. This is justifiable under some ethical perspectives but unjustifiable under other ethical perspectives. As Dreher and Sackett have noted, the utility of the outcome depends on the ethical perspective taken, and any single outcome will be not equally serve the interests of all stakeholders (e.g., individuals, organizations, or society).

Implications for Psychological Testing

The hypothetical situations described above highlight the differences between approaches to establishing the scientific validity of personnel selection tests and the implications that taking any particular approach has for the information a test will provide and the predictive utility of test outcomes. How a test is validated will also go a long way toward determining how to ethically and legally meet the legitimate needs of an organization to staff effectively. While psychology deals with the implication of scientific methods to solve problems and seek truth, when psychology ventures into business and industry, the goals became more complicated. Psychologists are faced with HR management’s needs to select effective and efficient workers within legal parameters and in a manner that assures employees that there will be a balance between business and human needs. The historical movement to use scientific methods to refine the efficiency of business and industry comes face to face with the balancing forces of government with its goal of protecting the individual.

These competing goals were reconciled in Title VII of the Civil Rights Act of 1964, which outlawed disparate treatment of employees and job seekers, and in Office of Federal Contract Compliance Programs regulations based on President Lyndon Johnson’s 1966 Executive Order No. 11246, which requires all federal contractors to implement written affirmative action programs that include specific goals and timetables for hiring minorities. These and subsequent federal rules demand that selection specialists design reliable and valid instruments that meet managers’ and owners’ needs for compliant personnel selection.

After 1964, personnel decision makers’ ability to use selection tests, including powerful tools such as cognitive tests, was limited by injunctions to avoid adverse impact. Herrnstein and Murray are probably the most outspoken opponents of the shackling of testing. These authors hold that cognitive tests in particular are highly
useful tools in the search for employees who are likely to be highly productive and successful on the job or in educational endeavors. They maintain that science has done its job and identified instruments that are efficient at predicting job success on an individual level. Herrnstein, a psychologist at Harvard, has argued that cognitive tests do what they were designed to accomplish, which is to separate out the most potentially successful from the least likely to succeed. Murray, a policy analyst and Bradley Fellow at the American Enterprise Institute, has pointed out that the clash between a laudable policy designed to mitigate disparate impact and an equally laudable scientific goal of developing and using valid tests is a complex issue that requires considerable expertise to evaluate.

Herrnstein and Murray have pointed out that science has reasonably definitively demonstrated that

- Different racial groups perform differently on intelligence tests, particularly IQ tests.
- Intelligence tests have considerable predictive validity when it comes to employees' future productivity and job success.
- Differences on mean and distribution of IQ can be used as efficiency enhancement tools rather than facilitators of racial discrimination.
- Ignoring real predictive differences among groups and the utility of those differences when selecting individuals is unfair, unscientific, unnecessary, economically absurd, and does not really aid protected groups in the long run.

They suggest that the data show that the plight of racially protected groups was changing for the positive prior to the enactment of the Civil Rights Act of 1964, the issuance of the *Griggs v. Duke Power* decision, and the publication of the *Uniform Guidelines*. In Herrnstein and Murray's opinion, affirmative action programs made small and early differences, but had no substantial positive influence on the plight of protected groups.56

Herrnstein and Murray recommend that the choices to deal with these problems are to

- Create tests that are legal under the current requirements.
- Choose among applicants with equal education, recognizing that affirmative action has produced differences ranging from 1.4 to 1.6 standard deviations in IQ within college graduates with common majors.
- Use race norming with top-down hiring methods.
- Return to the original concept of affirmative action by scrapping discrimination law and encouraging companies to do the right thing and exercise social conscience and fairness.57

In summary, the application of psychological testing to personnel selection has been problematic and presents both researchers and practitioners with difficult challenges. The controversy over the use of psychological tests and how to deal with...
adverse impact resulting from group differences continue with no definitive answers. Part III will describe some of the steps being taken to deal with the challenges created by psychological testing and the implications of the use of psychological tests, particularly some of the solutions suggested by newer theories and instruments of personality, for practitioners.

Notes


8 Ibid.

9 Ibid.; 3-10.


51 Cleary, T A. (1968), op cit.


56 Ibid.

57 Ibid.
Authors

Wesley A. Scroggins, PhD
Assistant Professor of Management
Department of Management
Missouri State University
901 S. National
Springfield, MO 65897
(417) 836-5505
wesscroggins@missouristate.edu

Steven L. Thomas, PhD
Professor of Management
Missouri State University
901 S. National
Springfield, MO 65897
(417) 836-5076
steventhomas@missouristate.edu

Jerry A. Morris, PsyD
Consultant
Morris & Morris, Inc.
815 S. Ash
Nevada, MO 64772
(417) 667-8352
cmhcjerry@sbcglobal.net

Wesley A. Scroggins, PhD, is an assistant professor of management at Missouri State University. He earned his PhD from New Mexico State University and has published several articles in the HR management areas of selection and compensation.

Steven L. Thomas, PhD, is a professor of management at Missouri State University. He earned his PhD from the University of Kansas and has published extensively in labor relations and HR management.

Jerry A. Morris, PsyD, has an MBA from Missouri State University and is a licensed psychologist, director of a mental health care center, author, and consultant.