THE BLACKWELL ENCYCLOPEDIA OF MANAGEMENT

SECOND EDITION

HUMAN RESOURCE MANAGEMENT

Edited by
Susan Cartwright
Manchester Business School,
University of Manchester

First edition edited by
Lawrence H. Peters, Charles R. Greer,
and Stuart A. Youngblood

Blackwell Publishing
2005

construct validity

Linda S. Gottfredson

Many psychological assessments purport to measure particular hypothetical traits (*see* TRAIT) or constructs such as extraversion, intelligence, self-efficacy, anxiety, or morale. Construct validity is a judgment about the extent to which an assessment actually measures the proposed trait in the populations of interest, and thus what can be appropriately inferred from individuals' scores on it. Validity is never a blanket judgment, but is established for specified uses of the assessment.

Construct Validation as Theory Testing

A construct is a tentative theory about an unobservable, underlying trait that is invoked to explain patterns of responses on an assessment. Construct validity requires evidence that individuals with different scores on the assessment actually behave as predicted by the theory. No single procedure or piece of evidence suffices to establish construct validity. The more evidence collected concerning the nature, causes, correlates, and effects of the attribute being measured, the clearer the inferences that may properly be drawn from scores for it.

Construct validation is a complex inferential process drawing on many sorts of evidence, i.e., a process of theory testing. Cronbach (1990: 183) distinguished between weak validation (an undirected, inductive process) and strong validation (a tough-minded testing of specific hypotheses). Positive evidence supports both the measure and the theory. Negative evidence means that the measure, the theory, or both may be faulty. Different procedures in this inferential process are often identified as different forms of validity. Although construct validity is sometimes treated as only one among various forms of validity, it is increasingly viewed as a concept unifying all of them (Messick, 1989).

Kinds of Evidence

Content-related evidence of validity (what used to be called CONTENT VALIDITY) refers to the care with which test items were chosen to represent the specific processes or content thought to instantiate the construct. For achievement tests, this means appropriate breadth and depth in sampling from the intended achievement domain, say, physics. For ability tests, it means sampling the mental processes thought to comprise the ability, such as visualizing objects in three-dimensional space for spatial ability. Content-related evidence is obtained by examining how the test items were developed or by having examinees report their thought processes while tackling items on the test. Such evidence increases the likelihood that an assessment will measure the intended construct, but it provides no proof it succeeds in doing so.

Criterion-related evidence of validity (CRITERION-RELATED VALIDITY) refers to the degree to which scores on the assessment correlate with other traits, behaviors, and outcomes (the criteria). We might ask, for example, how well students' IQ scores correlate with their current academic performance (CONCURRENT VALIDITY) or later JOB PERFORMANCE (PREDICTIVE VALIDITY).

Convergent and discriminant validity refer to the patterns of correlations predicted by the broader theory in which the construct is embedded. Two assessments that supposedly measure the same construct (e.g., intelligence) should correlate highly with each other and also with behaviors the theory says will be affected by the trait (e.g., performance in school or job training). Measures of the construct should correlate only weakly, however, with measures of different constructs (e.g., anxiety or creativity) or with supposedly unaffected outcomes (e.g., athletic prowess).

Evidence of DIFFERENTIAL VALIDITY and SINGLE-GROUP VALIDITY also affects the interpretation of test scores. For instance, there was once much concern that job aptitude tests predict job performance for white job applicants better than (or only) for ethnic minorities, and thus should not be used to infer the job qualifications of the latter. This issue was settled by meta-analyzing many small studies (Hunter, Schmidt, and Hunter, 1979).

Other research strategies are also useful in determining just what constructs different assessments are capturing, whatever the original intent. For example, the structure, relatedness, and homogeneity of the traits being measured can be clarified through factor analysis, both exploratory and confirmatory. If the construct is a developmental one, then longitudinal or cross-sectional studies should reveal predictable age differences. Scores should also differ, or not differ, for other subgroups or circumstances (e.g., gender, personality type, job tenure) in the manner predicted. Interventions to change traits can also test assumptions about them. For example, studies of adoption and compensatory education forced some rethinking about the malleability of intelligence and the consequent meaning of high versus low IQ scores.

Bibliography

Braun, H. I., Jackson, D. N., and Wiley, D. E. (2002). *The Role of Constructs in Psychological and Educational Measurement*. Mahwah, NJ: Erlbaum.

Cronbach, L. J. (1990). *Essentials of Psychological Testing*, 5th edn. New York: HarperCollins.

Cronbach, L. J. and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.

Hunter, J. E., Schmidt, F. L., and Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin*, 86, 721–35.

Messick, S. (1989). Validity of psychological measurement. In R. L. Linn (ed.), *Educational Measurement*, 3rd edn. New York: Macmillan, pp. 13–103.

Linda S. Gottfredson

A score on a standardized test is interpreted by comparing it to some external standard. When scores are compared to those of some reference population, they are called norm-referenced; when compared to some absolute performance standard, they are criterion-referenced. Norms are the distributions of scores (means, standard deviations, etc.) for a test's various reference groups. Normed test scores are most commonly reported as percentile ranks or standard scores, such as *z*, T, or IQ scores. Age- and grade-equivalents are sometimes reported, especially for achievement tests in elementary school, but they have more technical disadvantages and are prone to misinterpretation. Latent TRAIT or "scaled" scores provide a new form of developmental norms that solve some but not all the interpretive problems of age- and grade-equivalents.

Norm groups (also called reference groups, normative samples, or standardization samples) may be national or local, and represent different age, grade, or social groups. Broad or narrow, however, they must be representative of the populations in question, clearly defined and described, and appropriate for their intended purposes. Intelligence testing compares scores of children of the same age (*see* INTELLIGENCE TESTS). Academic achievement tests typically compare the scores of children in the same grade and often from the same school or geographic area. An employer might compare the aptitude scores of job applicants to those of individuals hired at particular plants in the last five years. PERSONALITY TESTS and VOCATIONAL INTEREST INVENTORIES often provide separate norms for males and females for COUNSELING purposes. The US CIVIL RIGHTS ACT OF 1991 outlawed the use of scores normed separately by race, color, religion, sex, or national origin for purposes of selection or referral in employment.

Bibliography

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Cronbach, L. J. (1990). *Essentials of Psychological Testing*, 5th edn. New York: HarperCollins.

Kampaus, R. W. (1993). *Clinical Assessment of Children's Intelligence*. Boston: Allyn and Bacon.

Linda S. Gottfredson

An intelligence test is a series of standardized tasks for assessing general cognitive ability. The tasks may be diverse, including, for example, words, numbers, designs, pictures, and blocks. Tests that include more than one item type often arrange them in subtests such as vocabulary, information, block design, comprehension, arithmetic, and picture completion. Factor analyses show that, whatever their differences in manifest content, all IQ tests, mental test batteries, and parts thereof measure primarily a single common factor, called the general mental ability factor (*g*, for short; Carroll, 1993).

Intelligence tests therefore measure a highly general capability, which is reflected in higher-order thinking skills such as efficient learning, reasoning, problem solving, and abstract thinking. This is in contrast to aptitude and achievement tests. Aptitude tests target narrower abilities, such as verbal, mechanical, or spatial aptitude. Achievement tests assess knowledge of specific school curricula, such as reading, science, or history. Intelligence tests tend to require less specific, more generally available knowledge, sometimes only elementary concepts such as in/out or large/small. The distinctions among the three types of test are not always clear. Some aptitude and achievement tests function like intelligence tests when test takers have been equally exposed to the subject matter being tested.

Origins and Use

Alfred Binet and his colleague Théophile Simon constructed the first modern intelligence test, in

1905, in response to the French government's desire to develop diagnostic and instructional procedures for mentally retarded children. American psychologists developed the first group intelligence tests (called the Army Alpha and Army Beta tests) during World War I, in response to the Army's need to screen millions of recruits. The Army Alpha required examinees to read; the Army Beta did not.

Interest in mental testing grew rapidly after World War I, and both the federal government and military services in the US developed test batteries for large-scale screening of individuals for jobs. Many schools, colleges, and private employers likewise adopted some of the many new tests on the market for selecting and placing students and employees. Some of the group-administered tests (such as the SAT) are hours long, whereas others (such as the 12-minute, 50-item Wonderlic Personnel Test) are very short. The most widely used individually administered intelligence tests today are, for school-age children, the Wechsler Intelligence Scale for Children-IV (WISC-IV) and, for adults, the Wechsler Adult Intelligence Scale-III (WAIS-III). These IQ batteries are administered orally and most of their subtests are untimed.

The major uses of intelligence tests include clinical diagnosis of individuals' behavior or achievement problems, vocational and educational guidance, PERSONNEL SELECTION, and placement into different education and training programs. Good professional practice requires that test scores be supplemented with other information when high-stakes decisions are being made about individuals (e.g., assigning a child to a special education class).

Individual tests are administered by highly trained professionals who exercise judgment in gaining rapport, administering prompts, and scoring the quality of responses. Group tests can be administered by less-trained individuals because they allow no discretion in administration and scoring. The construction and use of intelligence tests are governed by professional standards, principally the STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTING (American Educational Research Association et al., 1999).

Trends

The construction of intelligence tests is increasingly guided by explicit theories of intelligence and new evidence on the structure of mental abilities (i.e., the relations between the general factor, *g*, and the narrower group factors, such as verbal and spatial ability). Multivariate confirmatory factor analysis is often used to evaluate a new test's CONSTRUCT VALIDITY and whether its results are equally construct valid in different race, age, and gender groups (Keith, 1997). Theories on the biological basis of intelligence may someday lead to very different sorts of intelligence tests. For example, the last two decades have produced much research testing the notion that differences in intelligence originate primarily in differences in the speed and efficiency of brain processes (Deary, 2000). A wide variety of structural and physiological features of the brain (such as brain volume, rate of glucose metabolism, latency and shape of brain waves), as well as speed of perceiving exceedingly simple perceptual stimuli (inspection and reaction-time tasks), have been shown to correlate moderately with IQ when considered individually and sometimes strongly when measures are aggregated.

Legal and Social Issues

Test use has risen and fallen during the last century, depending on social and legal currents of the time (Wigdor and Garner, 1982). Public concern has focused on test fairness, because mental tests are often used in ways that affect people's lives. Selection and placement are two such uses. Although often warranted by the tests' predictive value, such uses make tests the focus of longstanding sociopolitical debates over equal opportunity.

Pervasive and sometimes large racial or ethnic disparities in test scores continue to fuel claims that intelligence tests are culturally biased. Extensive research (e.g., Jensen, 1980; Wigdor and Garner, 1982) has shown that they are not biased against native-born, English-speaking Americans, including blacks. Their use, however, often creates DISPARATE IMPACT, which has provoked much litigation. GRIGGS V. DUKE POWER, 401 US 424 (1971), *Larry P.* v. *Riles*,

495 F Supp. 926 (ND Cal., 1979), and similar court decisions have greatly affected the regulation and use of tests in employment and educational settings. Media reports of the foregoing issues have tended to misreport expert opinion on intelligence testing (Snyderman and Rothman, 1988).

Bibliography

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. New York: Cambridge University Press.

Cronbach, L. J. (1990). *Essentials of Psychological Testing*, 5th edn. New York: HarperCollins.

Deary, I. J. (2000). *Looking Down on Human Intelligence: From Psychometrics to the Brain*. Oxford: Oxford University Press.

Jensen, A. R. (1980). *Bias in Mental Testing*. New York: Free Press.

Keith, T. Z. (1997). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D. P. Flanagan, J. L. Genshaft, and P. L. Harrison (eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues*. New York: Guilford Press, pp. 373–402.

Sattler, J. M. (2001). *Assessment of Children: Cognitive Applications*, 4th edn. San Diego, CA: Jerome M. Sattler.

Snyderman, M. and Rothman, S. (1988). *The IQ Controversy, the Media and Public Policy*. New Brunswick, NJ: Transaction.

Wigdor, A. K. and Garner, W. R. (1982). *Ability Testing: Uses, Consequences, and Controversies. Part I: Report of the Committee*. Washington, DC: National Academy Press.