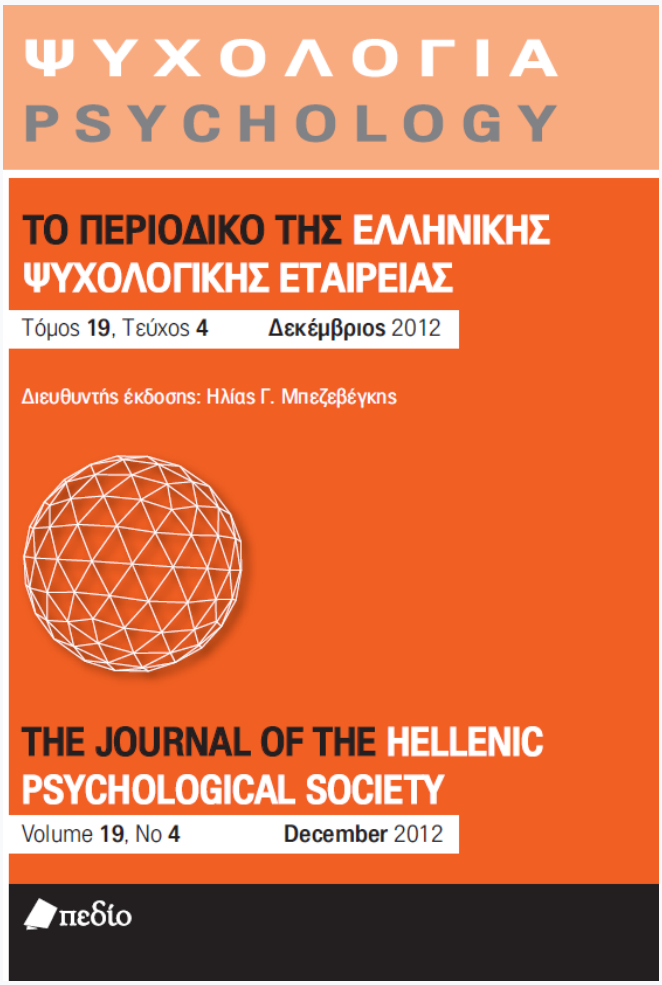
Psychology: the Journal of the Hellenic Psychological Society

Vol 19, No 4 (2012)



Test Development and Use Internationally

Thomas Oakland

doi: 10.12681/psy_hps.23702

To cite this article:

Oakland, T. (2020). Test Development and Use Internationally. *Psychology: The Journal of the Hellenic Psychological Society*, *19*(4), 472-480. https://doi.org/10.12681/psy_hps.23702

Test Development and Use Internationally[1]

THOMAS OAKLAND[2]

ABSTRACT

Test development and use constitute psychology's most important contribution to applied psychology. This article summarizes the status of test development and use internationally. Country and regional differences in test development and use exist and are explained in reference to qualities needed to support test development and use. The relevance of tests, like other technologies, should change in light of changing needs of those who rely on test data. Several somewhat prominent changes to which tests should respond are reviewed.

Key Words

Keywords: Test development internationally, Test use internationally, International Test Commission, Test use in Greece.

1. Introduction

Tests and other forms of assessment have been described as the flagship of applied psychology (Oakland, 2009). Their use in most countries may occur throughout the life span and be somewhat pervasive, often first occurring at birth with an Apgar measure as a simple and reliable method to quickly and summarily assess the health of newborn children. The use of tests may extend to an assessment of the quality of life together with cognitive and physical qualities that often decline in the elderly. Between birth and death, tests commonly are used to describe current behaviors, estimate future behaviors, provide guidance, help establish interventions, evaluate progress, screen for special needs, diagnose, place, and to credential, retain/promote. Test use may be most prominent in schools, with the ubiquitous teacher made tests, end-of-course examinations, together with high school graduation and college entrance exams being used commonly in education. Test use in research also is common and has contributed significantly to the establishment and growth of reliable and valid scholarship in psychology and other behavioral sciences. The empirical foundation for psychology is stronger in countries that have well developed testing resources.

---

1. Some content in this article was presented at the 13th Panhellenic Conference of Psychological Research, Athens, Greece, May 2011.

2. Address: University of Florida. 1921 SW 8th Drive, Gainesville, Fl 32601-8405. Tel.: (352) 376 8396. E-mail: Oakland@ufl.edu

2. Some Background Information

Greece's contributions to assessment have been considerable, beginning with the writings of Aristotle, Socrates, Plato, and other Greek philosophers that provided a foundation for understanding the nature of man. For example, the scholarly foundation for temperament and personality was established as early as 350 BC when Hippocrates (1923, 1994) described four humors or temperaments associated with body fluids thought to control or at least influence behavior. Later, Galen 1963; see also Kagan, 1994) extended Hippocrates' work by describing four pathological temperaments (i.e. choleric, melancholic, phlegmatic, and sanguine) derived from four bodily fluids. Thus, temperament's biological basis found its origin in these early Greek writings. Additionally, the Greeks pioneered efforts to diagnose mental retardation that relied on concepts of what we currently know to be adaptive behavior (Oakland & Harrison, 2008).

Psychological assessment is based on psychology's foundation principles of inter-individual and intra-individual differences. The importance of individual differences constitutes psychology's signature contribution to the behavioral sciences. Both research and practice are based on this concept. Psychological services are intended to describe and explain the nature of these differences, to examine their possible origins (e.g., as a function of gender, social-economic level, race, environmental conditions, and personal choices), and to trace their development (e.g., as a function of age). The applied practices of those engaged in clinical services acknowledge the importance of recognizing an individual's personal qualities and thus to provide one-of-a-kind services.

Test development and use reflect the importance of understanding individual differences. A goal of tests is to distinguish differences between and within persons. Tests that do not discriminate inter-individual and intra-individual differences hold little interest to psychologists.

3. Some Early History of Test Development and Use

The first recorded use of tests occurred in China approximately 3000 years ago with the establishment of a civil service examination of problem solving, visual spatial perception, divergent thinking, creativity, and knowledge of rights and rituals (Zhang, 1988). The People's Republic of China continues to hold national examinations each spring to help select government employees.

Many psychologists trace modern assessment to the efforts of three western pioneer psychologists: William Wundt in Leipzig Germany who established one of the first psychology laboratories in 1879, Frances Galton in London who established a laboratory to study individual differences ten years later, and Alfred Binet in Paris who established the first practical test of an individual's intelligence, first in 1904 and with revisions in 1905 and 1911. The 1911 version was adapted for use in many countries. The development of group tests of intelligence for use in World War I by the United States together with research on their reliability and validity helped establish respect for their use. The development of aptitude tests for use in World War II by the United States also contributed importantly to establishing respect for test use in society.

4. Status of Test Development and Use Internationally

The use of standardized tests is a common feature of society in developed countries. Test use occurs in three arenas: within schools, in industry, and through clinical services provided by practitioners. A country's largest consumers of standards tests tend to be schools, followed closely by industry and then clinical practitioners.

Test use with children and youth. A survey of test development and use with children and youth in 44 countries (Oakland, 1995, 2004) found that psychologists used measures of intelligence and

personality most often; surprisingly, measures of achievement were used rarely. A follow-up survey during 2011-2012 of the status of test development and use with children and youth in 80 countries identified the following ten tests as those used most widely: Wechsler Intelligence Scale for Children, Ravens Progressive Matrices, Bender Gestalt Test, Wechsler Preschool and Primary Scale of Intelligence, Kaufman Assessment Battery for Children, Child Behavior Checklist, Wide Range Achievement Test, Children's Memory Scale, Children's Apperception Test, and the Wechsler Individual Achievement Test. At first, foreign developed tests often are used more commonly than locally developed tests, especially in the smaller and less industrialized countries. They typically are translated into the local language. Later, a number of tests are adapted for local use. Estimates of reliability are more common than estimates of validity. Measures of personality typically often on theory and lack norms. Respondents from almost all countries reported the need for additional tests to assess mental retardation, learning disabilities, emotional and social maladjustment, physical impairment, gifted and talented students, and those with visual or auditory impairments.

Test use with adults. A survey of test development and use with adults in 29 countries (Muniz, Prieto, Almeida, & Bartram, 1999) also found foreign developed tests used more commonly than locally developed tests, especially in the smaller and less industrialized countries. The Wechsler intelligence scales for adults, Myers-Briggs Type Indicator, Minnesota Multiphasic Personality Inventory, and the NEO Personality Inventory were used most frequently.

5. Summary of the status of test development and use internationally

Test development and use are strongest in Australia, Canada, most Western European countries and the United States. These counties display strong beliefs in science, technology, individual differences, and meritocracy. Test development and use are lower yet emerging somewhat in Asia, Eastern Europe, and Latin America. Test development is lowest in the 22 Arab countries and the 54 African countries. Few tests are available in most of the 12 countries with the largest populations (i.e. in rank order: China, India, Indonesia, Pakistan, Bangladesh, Russia, Japan, Philippines, and Vietnam). An estimated 75% of the world's population reside in countries that have few locally developed standardized psychological tests.

6. Conditions Needed to Develop Tests

Psychologists and other professionals working in most emerging countries that lack tests see the need for additional tests. Four conditions typically must exist for test development and use to flourish.

The need for tests precedes test development and use. Test development occurs in response to professional and social needs. Professionals who are potential test consumers must see a need for tests and to be committed to their purchase and frequent use. Test consumers commonly include educators, counselors, management specialists, medical specialists, occupational therapists, physical therapists, psychologists, social workers, speech pathologists, and other professionals.

The public also constitutes test consumers and must see value in tests; to find their data to be reliable, valid, and maintained confidentially; and to pay for testing services. Professionals who use tests and those whose lives are impacted by test use must display positive attitudes toward science and technology in order to value test use.

Availability of educational institutions to properly prepare professionals. Professionals must receive suitable education and training at undergraduate and graduate levels to use tests properly. Thus, university-based undergraduate and graduate programs are needed to prepare specialists in test development and use. Professors also commonly take leadership for test

development or for research that examine a test's psychometric qualities.

This requires the availability of professors with strong academic and professional backgrounds in test development and use—a human resource requirement that may be in short supply in many emerging markets. For example, few specialists are prepared in psychometry—a specialization that focuses on test development. Countries that have relied on theory to guide psychology and minimized the value of empirically-based interventions, that have not properly kept pace with advances in statistics and psychometrics, and that instead have relied on a few widely used translated tests may lack the internal resources needed to foster test development and use.

Availability of a test industry infrastructure. A testing industry generally is needed to assist in test adaption, test development, marketing, and sales. The basis of this industry may be public (i.e., government-supported) or private. Test companies should employ professionals who display expertise in the various specializations needed to develop, adapt, and market tests. The market for a test must be sufficiently large to justify its development. Tests generally should have a commercial value. Test adaptation or development can be very costly. Those who invest their money, time, talents, and other resources can expect a return on their investments. Thus, tests must be purchased, not photocopied. In summary, a properly financed and managed infrastructure is needed to develop and market tests.

Available and committed professional associations Professional associations are needed to establish and maintain high standards for test development, test adaptation, and use; they also are needed to advocate for test use. These qualities constitute part of the desired testing infrastructure.

A profession's strength is directly seen in the strength of its national professional associations. These associations must be firmly and publically committed to the following four qualities that support its strong commitment to test development and use in psychology: value individual differences, be quantatively oriented, advocate for establishing and enforcing high professional standards (e.g., including, in part, to test development, adaptation, and use) and to advocate for the value of using tests to address important social and personal issues. This latter quality requires scholarship on the impact of test use in a country.

Available technical guidelines and standards The work of persons engaged in developing or adapting tests and using them can be guided by scholarship that addresses these issues, authored by scholars who have extensive academic and professional experiences. Guidelines and standards that impact tests address four separate yet interlocking issues: technical standards for developing tests, ethical standards for using them, clinical applications of tests, and legal standards governing test use.

Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999) provide the most authoritative technical guide to test development and use. The most authoritative technical guide to test adaptation is provided by the International Test Commission's test adaptation guidelines (www.intestcom.org).

Most countries do not have professional associations of psychologists and thus do not have nationally established ethics codes. Ethics codes in some countries (Leach & Oakland, 2007) that have professional associations emphasize broad and virtuous qualities (e.g., respect, responsibility) intended to have a pervasive impact on psychological practice and do not focus on more specific interests (e.g., clinical practice, testing, advertising). Only Latvia's ethics code explicitly addresses issues pertaining to test adaptation.

The American Psychological Association's *2002 Ethical Principles of Psychologists and Code of Conduct* (www.american psychological association/2002 ethical principles and code of conduct) provides the most detailed guide for the ethical uses of tests. The International Test

Commission's Guidelines for Test Use discusses the fair and ethical use of tests with the intent to provide an internationally agreed framework from which standards for training and test user competence and qualifications could be derived (www.intestcom.org).

Clinical guidelines for test use commonly are found in peer-reviewed journals, authoritative textbooks, and best practice documents from professional associations. For example, the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision* (1995, 2000), its *International Edition,* and the World Health Organization's *International Classification of Diseases and Related Health Problems, Tenth Edition* (1992) commonly guide the use of assessment data when making diagnoses. The International Test Commission's International Guidelines on Computer-Based and Internet-Delivered Testing highlight good practice issues in computer-based testing and testing delivered over the Internet (www.intestcom.org).

Legal standards governing test use, when available, generally take precedence over technical, ethical, and clinical guidelines. The legal basis for behaviors typically is defined at the national level, not the international level. Enacted laws, case laws, and administrative decisions govern the practice of psychology in all countries.

7. Social and Professional Changes to which Test Development and Use Should Respond

Testing practices should respond to current needs within a country as well as the current status of a profession. These needs change and thus testing practices need to change. The following discusses changes occurring in a number of countries.

Decrease emphasis on traditional forms of diagnosis. In many countries, clinical psychology first was dependent on and later emerged from psychiatry. Some older clinical psychologists may remember a time when clinical psychologists largely acquired data to assist psychiatrists in arriving at their diagnoses. Clinical psychologists generally retained their medical model emphasis on diagnosis after developing a practice independent of psychiatry.

Prevailing thought within psychology suggests that knowledge of a person's diagnosis does not identify needed interventions or the methods to promote development. The public wants and demands to know more than a diagnosis. Additionally, much of psychology has moved from a medical model to a social systems model for behavior.

Increase emphasis on describing behavior within the social and environmental contexts in which it occurs. Assessment methods need to adapt to a social systems model, one in which the process of assessing and interpreting test data is enhanced when it is viewed as a joint activity between the professional and those with whom the professional is working. Everyone engaged in the assessment process (e.g., the examiner, examinee, family members, educators) has a stake in the outcomes, including the interpretation of data and their implications, including desired interventions, if needed, and processes to attain them. This engagement may require the assessment of multiple traits through the use of multiple assessment method so as to acquire data from multiple reliable sources that describe behaviors in multiple settings and over multiple time periods.

The World Health Organization's International Classification of Functioning, Disability and Health (ICF; World Health Organization, 2001) provides a bio-psycho-social framework for viewing behaviors from three broad and different perspectives: health conditions, environmental factors, and personal factors. All interact, leading to the attainment of latent traits that have the potential to be displayed. The purpose of the ICF is to describe the complex interaction various qualities exert on health, including mental health. Its purpose is not to guide diagnoses. Practitioners can continue to rely on the WHO's International Classification of Diseases and Related Health Problems, Tenth Edition (1992) for this purpose if needed.

Emphasize interventions that either develop or rehabilitate behaviors. The public is demanding information that assists in promoting development. The term *empirically-supported interventions* increasingly is used in psychological literature to emphasize the need to acquire data that inform both the nature of interventions as well as the process that enhances the attainment of desired behaviors. Psychology's focus on general traits such as general intelligence, fine and gross motor skills, and adaptive behavior is inconsistent with these efforts. They do not meet the following criteria for effective interventions. Effectiveness increases when we focus on behaviors that are small and well defined, are modifiable, occur naturally in one's environment, are seen by our client and others as important, and the display of the newly learned behaviors is reinforced by one's self and others. In short, attempts to improve broad concepts and traits important to psychology have not been successful.

Acquire functional information: link data to interventions. The preferred theme, linking assessment and interventions, is apparent from the information discussed above. Psychologists need and want testing resources that enhance their ability to acquire functional information—that which can lead to a meaningful impact on one's life. They recognize the need to link assessment data to interventions. For example, some tests provide suggestions for ways to promote desired daily living skills that may be undeveloped or not well used. Their use allows test users to link assessment data to empirically supported interventions. Thus, those engaged in test development and evaluation should attempt to identify specific behaviors that may be modified, are influenced by age (i.e., that change), and have functional and clinical value. Software that links test data to behavioral interventions also should be provided, consistent with the validity of test data for various purposes.

Improve efficiency of test use. Testing can take many hours and thus may be costly. As a result, psychologists are being asked to reduce the amount of time needed to collect, score, and report data. This creates a professional dilemma: psychologist desire more information and deserve to be compensated for their work while the public may be unable or unwilling to pay. The solutions to this dilemma often are not apparent. Some remedies include the use of briefer tests, those that can be completed without an interview, may be administered and scored by a technician or electronically, and using computers to acquire, score, and interpret data.

Some advocate the use of developing and administering cognitive tests based on item response theory rather than classical assessment methods (Mpofu & Oakland, 2010). These methods allow test developers to use item difficulty data to build different forms of a test that are appropriate to each examinee. Thus, although the test includes items that differ in difficulty, each test is tailor made to include only those items that are most discriminating for the specific person being tested.

High stakes decisions require high stakes assessment methods. Test data can have different impacts on a person's life. Some test data have a low stakes impact (e.g., a score on a spelling test). Other test data have a high stakes impact (e.g., when a person is admitted, hired, or fired). Different standards are needed for low and high stakes test use. Lower stakes testing tolerates lower levels of reliability and may depend only on content validity. In contrast, higher stakes testing requires higher levels of reliability (e.g., above .90 when used individually) as well as well-supported validity evidence. These efforts may require the attainment and analysis of considerable clinical data—work that can increase tests construction costs considerably and thus justify higher test prices. High stakes assessments generally require the assessment of multiple traits through the use of multiple assessment method so as to acquire data from multiple reliable sources that describe behaviors in multiple settings and over multiple time periods. These conditions also increase costs.

Testing resources are needed in countries that currently lack the resources needed to

initiate their development. The availability of reliable and valid tests in a country will greatly enhance professional prestige as well as professional competence to deliver services that impact a range of decisions, including their use to describe current behaviors, estimate future behaviors, provide guidance, help establish interventions, evaluate progress, screen for special needs, diagnose, place, and to credential, retain/promote. Discussions among those engaged in test development will need to consider the cost-benefit ratio of adapting tests or devising new tests. Although those engaged in these discussions initially may prefer devising new tests rather than adapting well-established tests, models used in other countries suggest value in initially adapting tests while concurrently strengthening a county's infrastructure needed to assume a stronger role in test development.

Various models that promote test development exist (Oakland, 2009). Basic features of two test adaptation models are summarized. In one model, psychologists in four Eastern European countries worked with a test author who identified the best subtests from a widely used measure of intelligence in light of the target culture, provided the visual materials, and trained the personnel who collected the norm data on children from four countries. The test author assisted in data analysis, including the factor structure and norms tables. Using another model, a test development company in another Eastern European country obtained permission from test companies and test authors in the United States and elsewhere to adapt some the world's leading tests for use in its country. Two or more tests may be adapted concurrently, thus saving costs and providing validity data. These or other models may be relevant to initiate test adaptation activities in other emerging markets.

8. External and Internal Conditions that may Impact Test Development and Use

The development of psychology, including its technical features, is impacted by some conditions that are external to psychology—those over which the profession has less control, as well as some that are internal to psychology—those over which the profession has more control. Some external and internal conditions are reviewed below.

Some external conditions that impact test development and use. Test development and use generally are stronger in countries that have a sufficiently large population to warrant test use as well as sufficient wealth to pay for their use, including an economy that is stable, provides sufficient support for education, business and industry leaders see value in test use, test purchase is supported, and funds are used to establish and sustain a testing infrastructure.

The government and public should display attitudes that value tests to help address important social and personal problems, feel assured that data will remain secure and confidential, and, importantly, that individual differences and meritocracy are honored. In contrast, countries that are strongly socialistic, in which egalitarianism is the prevailing belief, will not support test development and use.

Some internal conditions within psychology that impact test development and use Test development and use are strong when psychologists value individual differences and meritocracy; value science, technology, and empiricism; help ensure that data are secure and remain confidential; value tests as helping to address important social and personal problems; identify ways that test use can help address important social and personal issue; and maintain high professional standards in their work.

Professionals need to purchase and use tests, not to photocopy them; to conduct research that examines tests' reliability, validity, and fairness; and to engage in test development efforts that address important social or personal needs as authors and collaborators, data collectors, and reviewers.

Additionally, one or more strong professional associations that advocate for test develop and use are needed, in part, to establish and enforce technical, ethical, and professional standards;

promote the need for undergraduate and graduate preparation programs; and help create and publish books, journals, and other scholarly publications that address testing issues. They also need to work closely with educational institutions to help ensure that suitable numbers of professionals are being prepared and that their preparation is consistent with current and emerging needs, standards, and guidelines.

References

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

American Psychiatric Association (1995). *Diagnostic and statistical manual of mental disorders, fourth edition (DSM-IV: International Version with ICD-10 Codes).* Washington DC: American Psychiatric Association.

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders, fourth edition, text revision*. Washington DC: American Psychiatric Association.

Hippocrates. (1923). *English and Greek Works* (W.H.S. Jones, Trans.). New York: Putnam.

Hippocrates. (1994). *Hippocrates* (R.H. Witherspon, Trans.). Cambridge, MA: Harvard University Press.

Kagan, J. (1994). *Galen's Prophecy: Temperament in Human Nature*. New York: Basic Books.

Leach, M.M., & Oakland, T. (2007). Ethics standards impacting test development and use: A review of 31 ethics codes impacting practices in 35 countries. *International Journal of Testing*, 7, 71-88.

Mpofu, E., & Oakland, T. (Eds.) (2010). *Assessment in Rehabilitation and Health*. Upper Saddle River, NJ: Merrill.

Muniz, J., Prieto, G., Almeida, L., & Bartram, D. (1999). Test use in Spain, Portugal and Latin American Countries. *European Journal of Psychological Assessment, 15,* 151-157.

Oakland, T. (1995). Test use with children and youth internationally: Current status and future directions. In T. Oakland & R. Hambleton (Eds.)*, International perspectives on academic assessment*. Boston, MA: Kluwer Academic Publishers.

Oakland, T. (2004). Use of educational and psychological tests internationally. *Applied Psychology: International Review, 53,* 157-172.

Oakland, T. (2009). How universal are test development and use? In E. Grigorenko (Ed.). Assessment of Abilities and Competencies in an Era of Globalization. New York: Springer, Publisher, 1-40.

Oakland, T, & Harrison, P, (2008) Adaptive Behaviors and Skills: An Introduction. In T. Oakland & P. Harrison (Eds.) *Adaptive Behavior Assessment System-II: Behavior Assessment System-II: Clinical use and interpretation.* San Diego, CA Elsevier, 3-20

World Health Organization (1992). *International statistical classification of diseases and related health problems, tenth revision* (ICD-10). Geneva, Switzerland, Author.

World Health Organization (2001). *International classification of functioning, disability and health* (ICF). Geneva, Switzerland: Author.

Zhang, H. (1988). Psychological measurement in China. *International Journal of Psychology, 23*, 101-117