# THE USE OF PSYCHOLOGICAL TESTS AND ANALYZING THE CONCEPT OFVALIDITY IN PSYCHOLOGICAL TESTING

**Tasnim Bibi Kazi**
*Academic and Lecturer at the Regent Business School, Durban, South Africa*

.

**ABSTRACT**
This paper attempts to historically trace and asses the use of psychological tests and the value of using them. On the other hand the paper does not pretend to analyze every variable that encompasses this vast and complex field of study.
The history of testing can be traced back to 2200 B.C. when proficiency testing took place in China, and 4000 years ago when the Chinese used tests for civil selection (Cohen, Swerdlik and Smith, 1992; Friedenberg, 1995). As Shelly and Cohen (1986:3) put it: "Long before there were psychologists there were psychological tests" (p3). And as long as there were tests, there were debates and arguments surrounding it. In the 40s, Hoffman (1962: 7) points out, "it was manifestly useless to raise even a question about the value and effect of these tests" because it was believed that individuals could be properly evaluated if given a range of psychological tests (Cohen et al., 1992). In the 1960s this changed, and proponents of testing began "fighting… the irresistible force of the argument which says that their questions are in practice often bad and in theory very dangerous" (Hoffman, 1962:8). This essay will look at anti-test arguments and the proponents' responses in relation to: psychometric properties, social and cultural factors, privacy issues and the "correct" versus. the "best" answer.

***Key Words:*** *Psychological Validity, Testing, Statistics, Proponents, Prediction*

The strongest argument for proponents of psychological testing is the psychometric properties of the test. According to Hoffman (1962: 60), testers "do not hesitate to point out that they have statistics to prove [tests] valid and reliable… [And] seem to believe that their scientific routines place them in an impregnable position so far as outside criticism is concerned." Hoffman (1962: 135) from the side of the anti-test group, states that although statistics can be misleading and cannot defend all types of criticisms, proponents of testing believe "criticisms unbacked by specific statistics may be dismissed as mere opinions… [Because] the testers build their tests on a statistical foundation, and defend their tests statistically."

In 1913, when John B. Watson declared that psychology is about the prediction of behavior, psychologists started creating tests that would predict individuals' behavior and performance. Predictive validity, as Shelly and Cohen (1986:83) state "is perhaps at the heart of the matter."

During World War II, the United States Army Air Force used a battery of tests developed by psychologists for classification and placement. According to Scroggins, Thomas and Morris (2008a:104) "the capacity of psychological tests to find and predict merit was well documented by military psychologists… by the 1940s." However, Thorndike and Hagen's large-scale statistical study in 1959 showed otherwise (Hoffman, 1962). In their study, they looked at 17000 men given one version of the test in late 1943, and acquired follow-up information on 10000 of these men in 1955 and 1956. Using this information, they separated the men into occupational categories and estimated success using a range of criterion including income, feelings of success, and time in occupation. Over 12000 correlations were analyzed between tests and success criterion in 90 occupations. From this data, Thorndike and Hagen found no convincing evidence to "predict degree of success within occupation insofar as this is represented in [our] criterion measures… We should view the long-range prediction of occupational success by aptitude tests with a good deal of skepticism and take a very restrained view as to how much can be accomplished in this direction" (Hoffman, 1962:148).

Many studies (Dale, 2003; Gregory, 2007) Oakes, Ferris, Martocchio, Buckley and Broach, 2001; McCurley and Murphy, 2005; Bertua, Anderson and Salgado, (2005) have since disproved this assertion, providing strong estimates of predictive validity. On the other hand, Scroggins, Thomas and Morris (2008b: 188) state that "test items, regardless of predictive validity, will not be perceived to be fair and will not be accepted… if the items are perceived to be unrelated." This idea of relevance, also known as content or face validity has always been an argument for the anti-test movement. Lasson (1992: 35) mentions an instance when a man encounters the following question in a pre-employment test: "Have you ever changed price tags in a store because the prices were too high?" and found it to be "intrusive and irrelevant." This links to issues of privacy, dealt with further on. In 1975, Mendel (as cited in Drenth, Thierry and Wolf, (1998: 37) published a book on the legal protection of applicants, and stated that because employers raise expectations by providing a prospect of a job, "whatever happens during the selection process must be relevant and should take into account these expectations." He also adds that applicants are in "a relatively defenseless position and are… the most vulnerable party" (Drenth et al., 1998::37).

With respect to content validity, the above question was found in an honesty test. It will be easily argued that this is not a norm in psychological tests. The Stanford-Binet, Wechsler Adult Intelligence Scale (WAIS-III), and the Wechsler Intelligence Scale for Children (WISC-III), for example, all have evidence to support not only content validity, but construct and criterion-related validity as well (Murphy and Davidshofer, 2001). The Stanford-Binet is valid as a predictor of academic success, has high correlations with other well-validated ability tests, and tests a range of items on judgment and reasoning. The same cannot always be said of personality tests. Morgeson, Campion, Dipboye, Hollenbeck, Murphy and Schmitt (2007a/b), with reference to the personnel selection context, state that personality tests have low predictive validity, these have not changed much over time, content validity is low, and even when accounting for faking (which together with impression management is a major debate surrounding use of personality tests), validities remain unchanged. The authors suggest honesty and lack of insight should not be taken for granted and that personality tests be used with other

tests. Nevertheless, a recent study by Meyer et al. (2001:128; 135) reviewed data from more than 125 meta-analyses on test validity and conclude that: "psychological test validity is strong and compelling [and] psychological test validity is comparable to medical test validity," and state that the study clearly confirms "the very strong and positive evidence that already exists on the value of psychological tests." However, the authors, in a response to rebuttals and problems about their review from others (Smith, 2002; Garb, Klein and Grove, 2002; Fernandez-Ballesteros, 2002; Hunsley, 2002), make an important statement: "Validity coefficients alone do not tell the whole story about the merits of a test, but they appropriately serve as a central foundation" (Meyer et al., 2002: 141).

The reliability of tests depends on two things: internal consistency and test-retest stability. In test-retest stability, individuals should come up with similar scores each time they are tested. In internal consistency, test items judging similar behaviors or attitudes ought to get similar answers (Hoffman, 1962; Shelly and Cohen, 1986). The anti-test movement had much to say about reliability. Hoffman (1962) offers an example where a professor looked at two students who wrote an intelligence exam during November 1961 and again in January 1962. The first student raised his mark from the 29$^{th}$ to the 69$^{th}$ percentile, the other from the 61$^{st}$ to the 94$^{th}$. The professor concluded: "Either the students greatly benefit by repeating the examinations or that the scores are quite meaningless or both" (Hoffman, 1962: 57). Clearly, he was questioning test-retest stability. Shelly and Cohen (1986: 107) agree: "The manner in which subjects react to test items does affect the test-retest stability… People may get bored with a test or remember the answers from the last time." To illustrate their point, they look at IQ tests in different studies. One study had undergraduate students take intelligence tests for eight weeks, and found that scores kept getting higher and higher. But it was two other studies that caught the attention of the anti-test movement (Shelly and Cohen, 1986). The first tested 355 8-year olds every year for four years and found that their IQ scores were unstable, that is, it varied considerably. The second study administered Wechsler pre-school scales to 85 children and then administered the WISC-R to them eleven years later. This study found IQ to be stable with strong full-scale IQ correlations of 0.86. What the anti-test movement asked was this: "How does one explain the contradictory nature of these two studies? It is clearly an important issue because, in the maelstroms of testing, intelligence test scores have a reputation for being, at least, reliable" (Shelly and Cohen, 1986: 113).

But, current and recent studies are different (Murphy and Davidshofer, 2001; McCurley & Murphy, 2005). The Stanford-Binet very rarely shows low levels of reliability, regardless of age or IQ and has an internal consistency of 0.88. The WAIS-III has given consistent evidence to support reliability of all its editions, with a split-half reliability exceeding 0.95, and verbal and performance IQ reliabilities between 0.90 and 0.95. The WISC-III is also highly reliable, with test-retest reliabilities exceeding 0.90, and individual subtest reliabilities of 0.70 or higher. With regard to personality tests, Miles, Shevlin and McGhee (1999: 147) did a study on whether the reliability of the Eysenck Personality Questionnaire (EPQ) differed across gender: "This… may occur if the meanings of any questions could be interpreted differently according to cultural context or expectations… Psychologists may [then] find themselves using a test that is not appropriate for a proportion of the people they are testing. Lower reliability may lead to

misclassification." The results of this study showed no significant differences between males and females across all four subscales of the EPQ.

What is interesting about the above reliability study is that it looks at social and cultural factors surrounding test usage. Often overlapping with these factors is the issue of privacy, mentioned above. Together these issues provide the anti-test movements with strong criticisms for the rejection of psychological tests. Many psychological tests inquire about topics that are considered to be sensitive, personal and private, and are therefore "widely regarded as unwarranted invasions of privacy" (Murphy and Davidshofer, 2001: 57). Currently some tests, like the Minnesota Multiphasic Personality Inventory (MMPI), contain items dealing with sex, religion, bladder control, family relations, unusual thinking, and other sensitive topics. Murphy and Davidshofer (2001) point out that only responding to the test can cause unease, while the thought that answers could fall into the hands of an unscrupulous person can result in anxiety about invasion of privacy. Tests like these pose a greater threat to invasion of privacy than ability or achievement scores that are thought to be relevant because it is believed that decisions makers have a valid reason to know the scores.

Issues about the invasion of privacy therefore rise mainly in connection with personality tests, but may still apply to all tests. For example, in some tests, questions that reveal motivational, emotional and attitudinal traits are disguised, resulting in the individual revealing information without being aware. It is argued that it is necessary for individuals to be unaware of how responses are interpreted, but there is also the issue of being subjected to a test under false pre tense (Anastasi and Urbina, 1997). Anastasi and Urbina (1997: 540) state: "The fact that psychological tests are often singled out in discussions of invasions of privacy probably reflects prevalent misconceptions about tests as well as their frequent misuse as a sole basis for decisions about individuals." Thus, when the use of psychological tests flourished, so did the concern of the use of the data gained from tests. Cohen et al. (1992: 9) state that in late 1950s tests were increasingly being seen as 'tools': "In the hands of a capable person they can be used with remarkable outcomes. In the hands of a fool or an unscrupulous person they become pseudoscientific perversion."Thus, besides issues of validity and reliability, issues of privacy and bias were raised in courts, and "public scrutiny of psychological testing reached its zenith in 1965" (Cohen et al., 1992: 8). The anti- and pro-test debate was not only scientific, but also fuelled by sociocultural factors.

By the early 1970s, the anti-test movement was widespread. Psychological tests were banned in some schools, due to invasion of privacy and minority group bias (Irvine & Berry, (1988). One of the reasons was because tests placed students into categories such as: Average, Feebleminded, Non-English Immigrants, Gifted, Delinquents, and so on. Monahan (1998: 6) says the "false aura of scientific objectivity" fostered bias, and it was believed that "the psychological effects of negative student classification… often created self-fulfilling prophecies of failure." Murphy and Davidshofer (2001: 65) state that because important decisions result from tests scores, it has a large impact on society: "Proponents of testing believe that psychological tests are preferable to other methods of making these same decisions; but… that does not relieve test developers and tests users from… considering the societal consequences of

their actions." While some psychologists, like Boring, hold that intelligence, for example, is what intelligence tests measure, critics such as Eysenck and Kamin, hold that these tests do not purely measure intelligence, but rather different 'types' of knowledge (Arnold, Robertson and Cooper, 1991). For this reason, intelligence tests were criticized as being culturally biased, as it was believed that intellectual development is largely dependent on environment and cultural contexts, so children who are rightly intelligent can be labelled unfairly by a test that has in fact failed to test their 'type' of knowledge, and has been created using other so-called 'normal' qualities. Murphy and Davidshofer (2001) provide a current example of this. The WISC-R and the Stanford-Binet contains questions such as: "What should you do if a younger child hits you?" In some cultures, the "correct" answer may be different from the answer one would expect from an intelligent, white, middle-class child.

During the 1970s, besides in education, intelligence tests were unpopular in personnel selection, also due to cultural and ethnic bias. Schmitt and Noe (as cited in Arnold et al., 1991) had shown that ethnic minority groups turn out lower scores on cognitive tests than whites. Jensen (as cited in Kline, 2000) provides reasons that critics of testing give for this: bias in testers, poorer schooling, reactions of Blacks to being tested by Whites, difference in motivation between the two groups regarding tests, selection of test items favoured by Whites, and fear of being tested. However, proponents of testing have shown that there is no bias in this testing because when test fairness was correlated with job success for different groups, no evidence of bias was found (Kline, 2000). Furthermore, Jensen (as cited in Kline, 2000) and Kline (2000) point out the misleading notions of test bias that should be rejected. Firstly, to cite the mean scores between groups as evidence of test bias is 'absurd' as group differences need to be considered. Second, although some items on tests may be culture-bound and the source of bias, it is impossible to know which of these are biased, but when they are found, with the help of psychometric measures, they will be removed. Finally, if a test were standardised on one population it is not necessarily biased against another, as other evidence would be needed to determine this.

Proponents of testing also point out that tests do the opposite, acting as a precaution against favouritism, subjectivity and/or unreliable decisions especially when stereotype and prejudice distort interpersonal evaluations. In the 1970s, Gardner (as cited in Anastasi and Urbina, 1997: 549) commented on the use of tests in schools: "The tests couldn't see whether the youngster was in rags or in tweeds, and they couldn't hear the accents of the slum. The tests revealed intellectual gifts at every level of the population." Moreover, proponents argue that if tests were to be removed, decision makers will have to rely on the old methods of letters of recommendation, interviews and grade averages. Although these methods are still used, they are used in conjunction with tests, not in place of them. Wigdor and Gardner (as cited in Anastasi and Urbina, 1997: 550) argue that tests brought validity, reliability and subjectivity to and eliminated bias from these older methods, which are less accurate in predicting performance, and that attacks on psychological tests "fail to differentiate between the positive contributions of testing to fairness in decision making and the misuses of tests as shortcut substitutes for carefully considered judgment." Tests can be misused with minorities, but they can be misused with others as well. Anastasi and Urbina (1997: 550) state: "When evaluating the social consequences of testing, we need to assess carefully the social consequences of *not* testing and

thus having to rely on other procedures for decision making that are less uniformly fair than testing."

Finally, Hoffman (1962) extensively discusses another old argument of the anti-test movement: the arbitrariness in the determination of a "correct" answer, the "best" vs. the "correct" answer and ambiguity in test questions. In 1959, a concerned parent wrote to a newspaper about a question that appeared in his son's school entrance examination, which perfectly illustrates the above argument. The question was: "Which is the odd one out among cricket, football, billiards and hockey?" The next day, two letters appeared in the paper, the first stating that 'billiards' is the correct answer, the second, 'football', each giving a reason. The day after, an eminent philosopher wrote to the paper, giving sound reasons why each answer is the odd one out. The newspaper itself chose 'hockey'. A few days later, a teacher wrote to the paper, with answers and reasons, and concluded his letter with the following statement: "[I] wonder how far this question and questions of a similar nature is a true and reliable guide in the testing of intelligence" (Hoffman, 1962: 20). The topic had lost its fun, had become too serious, and was promptly dropped from the paper.

Hoffman (1962) argues that while most questions asked in psychological tests are simple and clear, ambiguity often infiltrates. Testers, when told an item is ambiguous, argue that individuals should judge answers from context, thereby confirming the presence of ambiguity. Rivlin (as cited in Hoffman, 1962: 73) verifies this: "There is risk that more than one answer can be defended as the best one." All testers subtly hint at ambiguity. For example, to date, in many tests individuals are asked to pick the "best" answer, or one that "best fits, instead of the "correct" answer. Moreover, with psychological tests, the subject-matter expert does not determine the "best" answer – statistics does. Reliance on statistics is emphasized in the following statement by a pro-test lobbyist: "Whether a particular item is or is not correct, valid, or illogical is irrelevant" (Hoffman, 1962: 82), as it is statistics that will determine whether or not an item should appear in a test. Thus, arguing that an answer is not "correct" is futile. Irrespective of the likelihood that it may not be "correct", statistics determines the "best" answer, that is, the one that will be given the mark (Hoffman, 1962).

At present, Cliff and Keats (2003: 24) believe that there has been a substantial decrease in the anti-test movement "as people realize that many of the objections arise from poor test practice and can be removed by proper choices of tests and adequate training of testers." Therefore, despite the compelling arguments of the anti-test movement for the rejection of psychological tests, these arguments are predominantly historical. I hold that psychological tests, with all its shortcomings and flaws, have valuable functions and play important roles, and so cannot be rejected. As Anastasi and Urbina (1997) argue, it would be irrational and unmerited to wrongfully discard a tool that, although may never be without its weaknesses, still proves to be irreplaceable. Rather than focus on criticism, one should work on improving tests, because even though weaknesses cannot be eliminated, they can, at least, be reduced.

**Acknowledgements:**

The author thanks Professor Anis Mahomed Karodia for very useful inputs in the preparation of this article and for reading and editing the manuscript. Professor Karodia is a Senior Faculty Member and Researcher at the Regent Business School, Durban, South Africa

## References

Anastasi, A. & Urbina, S. (1997). *Psychological testing*. 7th Edition. Upper Saddle River, NJ: Prentice-Hall.

Arnold, J. Robertson. I. T. and Cooper, C.L. (1991). *Work Psychology: Understanding human behaviour in the workplace*. London: Pitman Publishing.

Bertua, C., Anderson, N. and Salgado, J. F. (2005). The predictive validity of cognitive ability tests: A UK meta-analysis. *Journal of Occupational and Organizational Psychology, 78*, 387-409.

Cliff, N. and Keats, J.A. (2003). *Ordinal measurement in the behavioral sciences*. USA: Lawrence Erlbaum.

Cohen, R.J., Swerdlik, M.E. and Smith, D.K. (1992). *Psychological testing and assessment: An introduction to tests and measurement*. 2nd Edition. California: Mayfield Publishing Company.

Dale, M. (2003). *A Manager's Guide to Recruitment and Selection*. 2nd Edition. London: Kogan-Page.

Drenth, P.J.D., Thierry, H. & Wolff, C.J. (1998). *Handbook of Work and Organizational Psychology: Personnel Psychology*. USA: Psychology Press.

Fernandez-Ballesteros, R. (2002). How should psychological assessment be considered? *American Psychologist, 57*, 138-139.

Friedenburg, L. (1995). *Psychological testing: Design, analysis and use*. Masachussets: Allyn & Bacon.

Garb, H.N., Klein, D.F. & Grove, W.M. (2002). Comparison of medical and psychological tests. *American Psychologist, 57*, 137-138.

Gregory, R.J. (2007). *Psychological Testing: History, principles and applications*. 5th Edition. Boston: Pearson.

Hoffman, B. 1962. (2003). *The Tyranny of Testing*. New York: Dover Publications.

Hunsley, J. (2002). Psychological testing and psychological assessment: A closer examination. *American Psychologist, 57,* 139-140.

Irvine, S.H. & Berry, J.W. (1988). *Human Abilities in Cultural Context.* UK: Cambridge University Press.

Kline, P. (2000). *The Handbook of Psychological Testing*. 2nd Edition. London: Routledge.

Lasson, E.D. (1992). How good are integrity tests? *Personnel Journal, 71,* 35-37.

McCurley, M.J. & Murphy, K.J. (2005, August) *Psychological Testing*. Presented at the 31st Annual Advanced Family Law Course, Dallas.

Meyer, G.J., Finn, S.E., Eyde, L.D., Kay, G.G., Moreland, K.L., Dies, R.R., Eisman, E.J., Kubiszyn, T.W. & Reed, G.M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56*, 128-165.

Meyer, G.J., Finn, S.E., Eyde, L.D., Kay, G.G., Dies, R.R., Eisman, E.J., Kubiszyn, T.W. & Reed, G.M. (2002). Amplifying issues related to psychological testing and assessment. *American Psychologist, 57,* 140-141.

Miles, J.N.V., Shevlin, M. & McGhee, P.C. (1999). Gender differences in the reliability of the EPQ? A bootstrapping approach. *British Journal of Psychology, 90,* 145-154.

Monahan, T. (1998). The rise of standardized educational testing in the U.S.: A bibliographic overview. Retrieved 13 April 2009, from http://torinmonahan.com/papers/testing.pdf

Morgeson, F.P., Campion, M.A., Dipboye, R.L., Hollenbeck, J.R., Murphy, K. and Schmitt, N. (2007a). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60,* 683-729.

Morgeson, F.P., Campion, M.A., Dipboye, R.L., Hollenbeck, J.R., Murphy, K. and Schmitt, N. (2007b). Are we getting fooled again? Coming to terms with limitations in the use of personality tests for personnel selection. *Personnel Psychology, 60,* 1029-1049.

Murphy, K.R. & Davidshofer, C.O. (2001). *Psychological Testing: Principles and Applications.* 5th Edition. New Jersey: Prentice Hall.

Oakes, D.W., Ferris, G.R., Martocchio, J.J., Buckley, M.R. and Broach, D. (2001). Cognitive ability and personality predictors of training program skill acquisition and job performance. *Journal of Business and Psychology, 15,* 523-548.

Scroggins, W. A., Thomas, S.L. & Morris, J.A. (2008a). Psychological testing in personnel selection, Part I: A century of psychological testing. *Public Personnel Management, 37,* 99-109.

Scroggins, W. A., Thomas, S.L. & Morris, J.A. (2008b). Psychological testing in personnel selection, Part II: The refinement of methods and standards in employee selection. *Public Personnel Management, 37,* 185-198.

Shelley, D. & Cohen, D. (1986). *Testing Psychological Tests.* Great Britain: Billing & Sons.

Smith, D.A. (2002). Validity and values: Monetary and otherwise. *American Psychologist, 57,* 136-139.