Bias in personnel selection and occupational assessments: Theory and techniques for identifying and solving bias

Herbert Kanengoni

University of Fort Hare, Department of Industrial Psychology, P. Bag X1314, Alice, 5700, Republic of South Africa.

Accepted 25 April, 2013

The controversy over the use of occupational assessments and how to deal with adverse impact resulting from group differences continue with no definitive answers. Despite the fact that it is a scientific and precise tool for measurement, psychological testing has been accused of being a culturally biased procedure that results in discrimination against minority groups, particularly against job candidates and such discrimination and bias raise problems which have adverse legal implications in organisations. The concept of bias is crucial in personnel selection specifically in psychometric testing and is one of the major problems encountered by psychologists. Although most practitioners rely heavily on psychometric testing as the best selection method, it is important for those who rely on these tests to be very careful when using them to avoid bias. The paper focuses on the theory and techniques for identifying bias and improving the validity of occupational assessments.

Key words: Bias, personnel selection, occupational assessments, discrimination

INTRODUCTION

The application of psychometric testing to personnel selection has been problematic and presents both researchers and practitioners with difficult challenges. The controversy over the use of psychometric tests and how to deal with adverse impact resulting from group differences continue with no definitive answers. Despite the fact that it is a scientific and precise tool for measurement as noted by Harold (2003), psychological testing is a culturally biased procedure that results in discrimination against minority groups, particularly against job candidates and such discrimination and bias raise problems which have adverse legal implications. However, in order to curb such problems, there are techniques that can be implemented to identify and reduce bias in psychological testing.

According to Guion and Highhouse (2006), bias refers to systematic group differences in item responses, test scores or other assessments for reasons unrelated to the trait being assessed. Bias is systematic in the sense that it creates distortions for members of a particular group. It is a concept that is defined in terms of groups of examinees. For example, if a test item requires knowledge common in one group but not in another, and if that knowledge is relevant to the trait, then the item is biased, in addition to that, if that knowledge is irrelevant to the trait, then the item is also biased. Bias may occur because of the difference in culture; thus when an acceptable response depends on skills or information common in one culture but not in another. This is usually manifested in multi-national companies across countries.
Unfair discrimination is also related to bias. Guion and Highhouse (2006) further noted that discrimination occurs when persons with equal probabilities of success on the job have unequal probabilities of being hired for the job.

**Explaining the concept of bias**

From the definition provided above bias can be understood as follows: It is the meaning inferred from scores that may or may not be biased, not the test, per se; although intrinsic test characteristics may contribute to biased inferences. It is group related; the score of an individual test taker may be invalid, but bias is only one possible source of invalidity. Bias exists only if the instructions are presented so that many people in the group have a common misunderstanding and when the assessors believe that the groups of people being compared are equal with respect to the trait being measured; hence a measure of bias that does not disentangle itself from genuine group differences is not interruptible. Emphasis is on sources of group variances not on group means because sources of variances are easily identifiable. Also bias exists when other sources of variance influence scores in one group but not in another. With that in mind it should also be taken into cognizance that, bias is differential validity of a given interpretation of a test score for any definable, relevant subgroup of test takers (Differential Psychometric Validity). To investigate bias from this perspective therefore requires more than merely comparisons of correlation coefficients. Generally, the methods that can be used to analyse the bias in tests include simple factor analysis, simultaneous factor analyses, analysis of variances, and the Item Response Theory models particularly at the item level. Analysis of this bias tries to identify possible instances of unfair discrimination. When talking of bias, it is rather unfair to ignore the concept of discrimination.

**Discrimination**

According to the Code of Good practice (2006), a fair test should not discriminate an individual on the grounds of gender, race, colour, religion and national origin. Once discrimination is unfair, it becomes illegal; therefore any selection procedure such as a test or an interview with the effect of unfairly discriminating against people in a protected group is discrimination under the law, even if inadvertent. However, in some parts of the world, socio-economic groups, groups defined by cultural or intellectual habits and other kinds of groups without legal or political protection may be discriminated against with no threat of litigation; but in the broader sense of business, this can rob the organization of people with excellent qualifications. Therefore, for organisations to get rid of such discrimination it is better to focus on valid, job-related assessment which can reduce such instances of bias decisions.

As posited by Murphy (1989), in order to identify and address bias in tests, it is important to examine group differences in score distributions, in validity or in predictions and this can be achieved through considering distributional differences in mean scores, variances, skewness and psychometric differences that influence bias. Distributions may differ in variance across different groups, with the protected groups including people from disadvantaged, dysfunctional, poor backgrounds and also people with more education or higher socio-economic heritage. The distributional differences may stem from systematic sources on measurement error related to group membership when groups are defined or influenced by unmeasured third variable such as test-taking habits (Aguinis, 2004).

**PSYCHOMETRIC TESTS AND THEIR IMPORTANCE TO ORGANISATIONS**

Goliath (2008) argued that 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. In a bid to clarify this concept, Goliath (2008) further indicated that, psychometric testing has become a powerful and institutionalised tool that is broadly applied. Universities and trade schools, managers, engineers, psychologists, and government officials recognise the importance of using scientific methods and tools to manage human capital, the economics of business, and the national defense. Industrial and organisational psychologists have emerged as the prominent players in the development and dissemination of the tests and the evaluation of the tests.

Occupational assessments can be defined as psychological instruments and procedures developed and used by professionals in organisations for the purposes of making inferences about people in the workplace or those who want to join the workplace (McIntire and Miller, 2007). An organisation may use the results of the tests for selection purposes, promotional purposes or future development of job incumbents. Hartigan and Wigdor (1990) indicated that, 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 organisation.

**TYPES OF BIAS**

As highlighted earlier in this article, test content can be biased, bias can be cultural in nature, and it can take the
form of gender bias, criterion bias, and also item bias which can be found in some of the items that make up a test.

Test bias

In general terms, test bias is defined as invalidity or systematic error in how a test measures members of a particular group (Wicherts, 2007). Test bias is often an issue in the study of racial and ethnic group differences and gender differences. However, bias in how a test measures can also be an issue for subgroups identified by social classes, age, region, urban–rural environment. Test bias produces scores with systematically different meanings for people who are alike on the characteristic being measured. The interpretation of test scores is biased for or against members of a group if groups of people matched on the trait measured have different scores because of one or more sources of variance related to group membership. It is imperative to note that, bias is group related, thus, the score of an individual test taker may be invalid, but bias is only one possible source of invalidity. In test bias, there is bias in measurement, then bias in prediction.

Bias in measurement

Bias in measurement occurs when the test makes systematic errors in measuring a particular characteristic or attribute; for example, many say that most IQ tests may well be valid for middle-class whites but not for blacks or other minorities. In interviews, which are a type of test, research (Guion and Highhouse, 2006) has shown that there is a bias in favour of good-looking applicants. When identifying bias in measurement, methods used to determine construct validity may be adapted to determine whether a test measures the same thing or substantially different things from different groups in the population and to consider external correlates and internal evidence.

Bias in prediction

According to Guion and Highhouse (2006), bias in prediction occurs when the test makes systematic errors in predicting some outcome (or criterion). It is often suggested that tests used in academic admissions and in personnel selection under-predict the performance of minority applicants. In addition to that, a test may be useful for predicting the performance of one group; for example males but be less accurate in predicting the performance of females. Bias in prediction exists when the regression equations used to predict a relevant criterion for two or more separate groups differ. Intercepts, slopes; standard errors of estimate, systematic differences are known to exist in the average scores by whites and minorities.

Cultural bias

As pointed out by Harold (2003), although psychological tests are scientific and precise tools of measurement, they are culturally biased procedures that result in discrimination against minority groups, particularly against minority candidates. Cultural bias is compounded by other factors such as the item selection process, the content of the items, and the responses considered acceptable to those items. Cultural bias entails whether the psychological constructs have the same meaning from one culture to another; how the different items are interpreted by people from different cultures; actual content validity may be different for different cultures; does the psychological construct have the same meaning from one culture to another; how are the different items interpreted by people from different cultures; actual content validity may be different for different cultures.

Criterion bias

SIOPSA (2005), as cited by Coetzee and Schreuder (2012), posits that criterion bias is a systematic error resulting from criterion contamination or deficiency that differentially affects the criterion performance of different sub groups. According to Guion and Highhouse (2006), the criterion should be reliable, valid, and free from third variable biases; however, it should be noted that this is extremely difficult to accomplish. A serious attempt to assess criterion validity may in itself be a way of assessing criterion bias. Evidence of valid measurement of the criterion construct is the sort of evidence most appropriate.

Bias in individual items

This is referred to as differential item functioning (DIF). Some litigants have called tests biased merely because of group differences in item pass rates. For instance in the case of Golden Rule Insurance, Company, v. Washburn (1984), the settlement stipulated that, group item difficulties should differ by no more than 15. Researchers attempt to analyze bias or fairness especially where there is adverse impact. In order to identify bias, there should be evidence of differential validity, and professional judgments have to be made even with flawed data. When test scores seem to be biased, firstly, it is important to clarify needs.

ISSUES OF VALIDITY

Goliath (2008) indicated that, the assessment tools are
designed with the hope of facilitating optimal personnel development and utilisation of human capital. However, using scientific principles and quantitative methods to conserve and develop optimal human resources by forecasting behavioural tendencies over extended periods proved to be a considerable challenge. The challenge usually met lies with variable success, as the opportunities for error are great. Psychologists capture the testing agenda, but they sometimes struggle mightily to identify instruments that are accepted and widely judged to be useful. Kunnan and Shohamy (2000) noted that, 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 and free of bias 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. Guion and Highhouse (2006) point out that 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. However, for the sake of this paper more attention will be on content and construct validity. Content-related validation is a demonstration that the content of the test reflects important job-related behaviours 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.

In order to establish test validity, it is important to rely on validity generalisation, which is often used when organisations want to use a test for selecting individuals for a specific job and choose to use a test purchased from a commercial vendor. This generalisability refers to the necessity that validation procedures used must be consistent with accepted standards and that job analysis has indicated that the validity studies for the job relative to the instrument are substantially similar in terms of ability and work behaviour. When an instrument lacks such characteristics, one can then detect some form of bias in it. 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 validity of the test however compromises the issue of bias in some instances.

Maintaining validity and minimizing bias in tests

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. 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 be fair and acceptable. However, considering the multi-dimensional nature of job performance in the current era that has resulted in other approaches such as differential weighting selection components, the employer should consider available alternatives which achieve the legitimate business purpose with lesser adverse impact.

Goliath (2008) further posited that, modern personnel selection test choice or development should involve 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.

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. According to Goliath (2008), science has done its job and identified instruments that are efficient at predicting job success on an individual level. Herrnstein and Murray (1994) 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. However, there is a clash between a laudable policy designed to mitigate disparate impact and an equally laudable scientific goal of developing and using valid tests and this is a complex issue that requires considerable expertise to evaluate. From the discussion above, psychologists are faced with HR management's needs to select effective and efficient workers within legal parameters in a manner that assures employees that there will be a balance between business and human needs. This demands that selection specialists design reliable and valid instruments that meet managers’ and owners’ needs for compliant personnel selection.

THEORIES AND TECHNIQUES OF REDUCING BIAS IN PSYCHOMETRIC TESTS

Psychologists have come up with some techniques that can be used to deal with bias and adverse impact in psychological tests. In addition to that, there are some mechanisms which can be used but what Sackett and Divorce (2001) proposed are not effective and these include using differential item functioning (DIF) to identify and remove items functioning differently by group, providing coaching programs to improve scores for all, providing more generous time limits and alternating test
taking motivation.

**Differential item functioning**

Roth and Bobko (2000) are of the idea that differential item functioning is one of the strong methods of getting rid of bias since it determines whether items function comparably across groups by comparing item responses of group members with the same overall scores. According to Zumbo (1999) DIF statistical techniques are based on the principle that if different groups of test-takers (males and females) have roughly the same level of something (knowledge), and then they should perform similarly on individual test items regardless of group membership. In their essence, all DIF techniques match test takers from different groups according to their total test scores and then investigate how the different groups performed on individual test items to determine whether the test items are creating problems for a particular group.

**Expanding criterion and predictors**

Moreover, on techniques of reducing test bias, Sackett and Divorce (2001) report some support for expanding the criterion as a means of reducing subgroup differences thereby getting rid of bias. In addition to that, expanding the range of predictors is also of great importance because; supplementing the existing cognitive with additional predictors outside the cognitive domain can reduce the overall subgroup differences in some circumstances. Sackett and Divorce (2001) further indicated that, this strategy has received considerable attention because broadening the range of predictors has the potential to both reduce subgroup differences and increase validity. It should also be noted that adding predictors does not only reduce subgroup differences but also assists in predicting the more specific aspects of job performance.

**Banding**

Aguinis (2004) noted that, test score banding is one of the approaches that have emerged to deal with the problem of differential group performance on valid tests. It defines score ranges in which scores are treated as if he or she has the same score and likelihood for successfully performing the job.

**Differential prediction**

This involves accounting for the differences in slopes and intercepts between subgroups (Johnson et al., 2001) applied the logic of synthetic validity to pool data across jobs, thus making such analysis feasible in settings where samples within jobs are too small for adequate power. Sackett et al. (2003) showed that omitted variables that are correlated with both subgroup membership and the outcome of interest can bias attempts to estimate slope and intercept differences and offer strategies for addressing the omitted variables problems.

**Use of highly standardized tests**

It is advisable to use standardised tests because they are administered under uniform conditions; for example no matter where, when, by whom or to whom it is given, the test is administered in a similar way. The standardised tests are scored objectively, that is the procedures for scoring the test are specified in detail so that any number of trained scorers will arrive at the same score for the same set of responses. They are also designed to measure relative performance, for example they are not designed to measure absolute ability on a task. In order to measure relative performance, standardised tests are interpreted with reference to a comparable group of people. Practitioners should note that, although this method may take care of bias as far as test scoring is concerned, the possibilities of bias being built into the test remain.

**Implementing models of fairness**

There are many different perspectives from which to view fairness in personnel selection, and each has implications for psychological testing and decision making. Sackett et al. (2003) 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. In addition to that, models developed by Cleary and Thorndike, incorporate the issues of single-group validity versus differential validity and differential prediction.

**GUIDELINES AND SUGGESTIONS FOR AVOIDING CULTURAL BIAS IN TEST QUESTIONS**

**Readability of questions**

It is best to use clear, plain English. Manually check the
Readability indices found in major word processors to be sure they are working as intended (Wiesen, 2002).

**Reading Lists**

Often reading lists are a good idea because they give people the opportunity to prepare for the test. To allow for sufficient time to prepare, a reading list should best be published long before the examination date. All people should have an equal length of calendar time to study.

**Unnecessarily academic content**

When textbooks are used as source material for jobs without an academic requirement, there is a serious risk that they introduce problems akin to those discussed above concerning readability because text books are written merely for academic purpose than career.

**Every day practical reference material**

Sometimes it is possible to base test questions on local SOPs, ISO9000 documentation, local rules or laws or guidelines. These types of source material usually focus more on practical aspects of the job than are textbooks. The use of local rules, regulations, operating procedures, and guidelines is generally desirable to the extent that these are current and reflect what is or should be done on the job.

**Content which unnecessarily evokes emotional responses**

Some questions may arouse anxiety or other emotional response from some candidates. They may cause unequal testing conditions. For example, it would be best to avoid a table depicting incarceration rates by ethnic and religious group. Avoid sensitive topics (example, criminal justice system, HIV/Aids) unless they are germane to the subject of the test. Avoid using words or phrases with more than one meaning, avoid presenting age, gender, culture, or ethnic groups in a stereotyped fashion, even in wrong answers.

**Balanced or neutral representation as to age, gender, culture, and ethnic group**

One approach to this would have a balance of male and female names mentioned, rather than just one gender. If some people mentioned in the questions are in supervisory roles, they should not all be one gender or ethnic group; likewise for criminal roles, leadership roles, etc. One should strive for test content which puts all test takers at ease.

**Contamination with irrelevant test content**

Sometimes, questions which are intended to measure one area are also measuring another. Reading level (discussed above) may be the most common example of this, but there are others. For example, maths items which ask about import duties or increases or decreases in stock prices may confuse people who do not travel internationally or do not trade in stocks.

**Test Content which is equally familiar to all groups**

Most test areas can be measured with questions in a variety of sub-areas. Often this is not mentioned in the test outline. Sometimes, one or more areas are less familiar to some groups of candidates. For example, math questions which involve import duties or tariffs may be less familiar to people from lower socio-economic groups. As another example, mechanical aptitude questions which focus on the working or repair of cars may be less familiar to women.

**Context**

The context of the test questions or instructions should not be confusing. For example, if the instructions for a verbal analogy test include very dated words or concepts (example, inkwell and blotter), then test takers might think that every question should be evaluated in terms of possible dated meanings. This might have a greater negative effect on the less test wise candidates.

**Tricky questions**

Tricky questions should be avoided. Questions can be tricky for reasons of content, context, or wording. With respect to content, questions may be tricky due to: trivial nature of the key or the question, voluminous extraneous material in the question which goes beyond what is found on the job, multiple correct answers, or very fine distinctions between distracters and the key.

**Time limits**

According to Hartigan and Wigdor (1990), highly speeded tests should be undertaken with caution. Beyond trying to minimize the discomfort of candidates, there is some evidence that blacks are at a relative disadvantage in taking speeded tests.
Equal practice in test taking

Exposure to sample questions no doubt reduces anxiety. Beyond that, one study showed that people retested on an alternate version yielded a mean increase of .2 SDs for cognitive tests (Hartigan and Wigdor, 1990).

Equal access to pre-test coaching

Inequity in access to commercial coaching may result in test bias (perhaps only when evaluated against an external criterion, and so missed with the more often used internal measures of bias).

CONCLUSION

The concept of bias is crucial in personnel selection specifically in psychological testing, as discussed above; it is one of the major problems encountered by psychologists. However, as indicated in the paper, it is advisable to apply the techniques of trying to get rid of bias since this is the only way of managing the crisis. As much as psychological testing is regarded as the best selection method, it is important to those who rely on these tests to be very careful when using them to avoid the problems of bias.

REFERENCES
