THE VALIDITY OF A TEST BATTERY USED IN THE SELECTION OF APPRENTICE ELECTRICIANS

By

NIGEL ANDREW RITSON

Submitted in partial fulfilment of the requirements for the degree of:

MASTER OF ARTS

in the subject

INDUSTRIAL PSYCHOLOGY

at the

UNIVERSITY OF NATAL, DURBAN

SUPERVISOR: SONIA HILL

December 1999

DECLARATION OF ORIGINALITY

I hereby declare that this is the original work of the author unless specified in the text. This dissertation has not been submitted for a degree at any other university.

_____

Nigel Andrew Ritson
University of Natal, Durban

CONTENTS

Chapter Three

Chapter Four

Chapter Five

Chapter Six

Chapter Seven

List of tables

ACKNOWLEDGEMENTS

My sincere appreciation to all those who kept up their encouragement throughout.

The financial assistance of the Centre for Science Development towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at are those of the author and are not necessarily to be attributed to the Centre for Science Development.

Special thanks to my family, and most of all to Leigh-Ann, who was my never-ending source of motivation and support.

ABSTRACT

The purpose of this study is to assess the suitability of a psychometric test battery that is used by a large service department of the Durban City Council in the selection of apprentice electricians. The essence of the investigation is to determine whether the tests being used are appropriate for the purpose for which they are applied, as well as being justifiable within the current context in South Africa. A predictive validity analysis was carried out to determine whether the psychometric tests have a correlational relationship with the trade test which apprentices undergo.

The analysis was carried out on a sample of fifty five apprentices, consisting of a black group of 16 and a white group of 39. The results of the analysis were that the tests generally did not have a positive link with the results of the trade test. The only exception was the Mental Alertness test which showed a significant correlation with the trade test for the white sample only.

The value of the study lies in its usefulness to the organisation which uses the test battery in terms of providing a review of the effectiveness of the tests. Recommendations of the study are that alternative methods for apprentice selection be investigated.

CHAPTER ONE

INTRODUCTION

Psychometric assessment techniques have long been used in industry to assist in making a variety of decisions regarding employees, most notably in the area of selection. The rationale for using psychometric tests in the selection process lies in the purported ability of the testing instrument to accurately and objectively assess the applicant's ability to perform the work required in the job.

The guidelines compiled by the Society for Industrial Psychology of South Africa on the use of personnel selection procedures states that the "underlying assumption of any personnel selection procedure is that the procedures used can predict one or other important and relevant behavioural requirement or job performance aspect of the position" (Society for Industrial Psychology,1992:1. [Hereafter referred to as The Guidelines]).

It would therefore follow that any organisation using a psychometric assessment procedure in its selection of employees would be using such a method because it significantly assists in predicting whether the applicant possesses the behavioural requirements and competencies necessary to perform that job. On the basis of this an informed decision as to whether or not to employ that person can be made with some degree of accuracy.

A selection instrument should clearly only be used if the above statement is true. If there is any doubt regarding the ability of the test to provide an accurate idea of the applicant's future performance in the job, then the test itself should be analysed for suitability of purpose. This type of analysis should logically be carried out on a fairly regular basis to ensure that the testing instrument remains useful and accurate as conditions such as job requirements and applicant profile change.

Developments in South Africa since 1994 will ensure that the matter of psychometric testing and selection emerges as an increasingly contentious and debated point of concern. Pressure from labour is particularly imminent as unions are likely to become increasingly concerned that their members' interests are protected against discriminatory practises (Wheeler, 1993). It is envisaged that in future human resource practitioners will be increasingly required to defend their selection practises against accusations of being discriminatory (Taylor, 1987a; Veldsman, 1990). Furthermore, the onus will be placed on the practitioners to validate the tests they use and to give concrete empirical evidence that their selection practices are fair (The Guidelines, 1992). Psychometric tests will surely come under the greatest amount of scrutiny because of the suspicion that they may be biased against certain groups.

1.0 Background to the study

Durban Electricity, a large service department of the Durban Metropolitan Council, makes use of a variety of psychometric tests. Based on the principles mentioned, they select applicants on an annual basis for training as electrical apprentices.

The applicants are selected for training primarily on the basis of their performance on a number of psychometric tests. Applicants with a Std 8 (Grade 10) education are invited to apply for the twelve positions available. If successful, apprentices undergo training for a period of two and a half years. At the end of this period the apprentices are evaluated by means of an industry trade test which aims to assess their proficiency after training, and if successful on this test, become qualified electricians. The pass mark for the trade test is 50%.

It is obvious that there should be a strong link between candidate performance on the psychometric tests and performance on the trade test. If an applicant performs poorly on the psychometric tests he or she is not employed because the test user would infer that the probability that the candidate would also perform poorly on the trade test is strong. However, if a candidate performs adequately on the psychometric selection tests, the test user reasonably infers that he or she would achieve a passing result on the trade test. Based on the actual results obtained by apprentices on the trade test over the years, this hypothesis has become increasingly questioned.

Pieterse and Bowden (1993:24) state the following regarding the selection of apprentice electricians: "One electrical apprentice in three is likely to fail his trade test this year. Why is this happening? Why are young apprentices, presumably all psychometrically tested for suitability and with reasonable intelligence, failing - and failing badly?"

The November 1996 trade test results bear out this lament. Three out of fourteen apprentices (22% of candidates) failed their trade test. If this result is converted to a

pecuniary cost, the implications are alarming. Each apprentice is trained for an average of 30 months, at a salary of R1 500 and a training cost to the organisation of R2 000 per month. The cost to the organisation of training the three apprentices who failed is R105 000 each, giving a total cost R315 000 (figures provided by the Durban Electricity training centre). This is, in effect, the cost to the organisation of having three unsuccessful candidates. Understandably, the failure rate of apprentice electricians over the past few years has raised cause for concern, especially when one considers that based on their psychometric test results they should not have failed the trade test.

The concerns expressed above are not unique to Durban Electricity, and the same phenomenon has been identified by other organisations which train apprentices. Horn (1990), in assessing a psychometric test battery used by a major manufacturer of diesel engines expressed concern about the percentage of apprentices qualifying as artisans, and suggested that the test battery be re-assessed. The test battery in question included two of the psychometric tests used by Durban Electricity. Of the total sample examined in this study, 68% of the apprentices passed their trade test at Olifantsfontein in 1982; 86.8% in 1983 and 69.6% in 1984. The percentage of apprentices qualifying declined by 17.6% in one year alone. As a result, the selection process was rigorously scrutinised.

Essentially, the psychometric testing instruments are designed to filter out those who will not pass the trade test and identify those who will. If there is not a 100% pass rate on the trade test each year, then the magnifying glass should be brought to bear on the selection process.

There could of course be a reasonable explanation for an apprentice failing the trade test which are external to the selection procedure, for example nervousness, a recent traumatic

life experience, sickness, poor attendance during training, etc. However, with each year consistently producing a fairly large percentage of failures, the answer could lie deeper than this. It is the hypothesis of this research study that the explanation lies in the weakness of the initial selection process itself.

1.1 Purpose of the study

It is the aim of this study to investigate how effectively the psychometric test battery predicts applicants' performance after training, as measured by the trade test. The method used to achieve this will be to conduct a correlation analysis to determine the relationship between the two measurement criteria - the psychometric test scores and the trade test scores.

It will be possible to determine on the basis of this index of relationship whether the psychometric tests have a strong correspondence with trade test results, in other words answering the question: are the initial selection tests useful in predicting final after-training competence or not? In essence, the predictive validity of the psychometric tests for the context in which they are being applied will be investigated.

The value of such an investigation lies in providing the test user and other parties at Durban Electricity with an interest in apprentice selection with valuable and relevant

information regarding the predictive validity of the tests being used. The findings of this study could have further ramifications - the results obtained may either ratify the current procedure and resultingly give the test user more confidence in the appropriateness of the tests, or alternatively, if problems in the current procedure are identified, then these can be corrected or perhaps alternative selection techniques can be sought and applied in the future.

1.2 Outline of the investigation

In this chapter, the problem of the ability of the psychometric test battery to predict performance on the qualifying trade test was explained. The background to the study, the purpose thereof and the limitations were discussed.

In the following chapters, a background to the history of testing will be entered into in order to explain the origins of psychometric testing and the rationale behind the use of the specific tests used in the study. The intention of this historical overview is to describe how and why psychometric tests developed in order to understand the context in which we use them today, as well as to expose the possible limitations of psychological testing.

The psychometric tests which form the subject of the study will be described in detail in order to isolate the psychometric qualities and function of each. The trade test structure will then be explained. Overlapping qualities or competencies between the two assessment instruments will be highlighted, the purpose being to identify common

elements which will perhaps provide an explanation for the strength of correlation (or lack thereof) between them. Previous research findings will be discussed in order to provide support or offer suggestions for the improvement of the selection process.

A statistical analysis of the data will then be conducted, followed by a discussion of the findings.

CHAPTER TWO

2.0 Historical overview

In order to understand and to provide perspective on the rationale underlying the construction and choice of present day tests as well as to understand the central tenets of the psychometric tradition, an overview of the historical antecedents and origins of psychological testing shall be presented. This chapter will discuss the roots of psychological testing before discussing the properties of the psychometric tests used in the study, so as to provide a context to the weaknesses inherent in psychometric testing today, as well as to highlight the flaws inherent in the psychometric tests used in the battery for the selection of apprentice electricians.

Anastasi (1982) agrees that the direction in which contemporary psychological testing is progressing becomes clearer when considered in the light of the precursors of modern tests. Furthermore, the limitations as well as the advantages which characterise current tests become more obvious when viewed against the background from which they originated. To trace the origins of the historical development of psychological testing, one would have to regress in time back to the nineteenth century.

2.1 Early theories

Maloney and Ward (1976) broadly classify four general theoretical approaches to intelligence which vary in the degree of overlap. They are: (1) psychometric theories; (2) developmental theories; (3) learning theories and (4) neuro-biological theories. In

addition to the above is the information-processing approach which has come about during the last decade (Hughes, 1989).

2.1.1 Psychometric theories of intelligence

Psychometric theories of intelligence share the basic assumptions that individuals differ in ability (Hughes, 1989). The individual differences, consisting of attributes and traits, are quantifiable so that differences and similarities can be measured and expressed numerically.

Verster (1987) proposes that the most significant starting point is marked by Herbert Spencer (1855) - a biologist, physiologist, and philosopher who developed a profound interest in the then emerging notion of evolution, and published on this subject even before the appearance of Charles Darwin's book *The Origin of the Species.*

Spencer was the first to use the term 'intelligence' and suggested that the mind, or intellect, could be described as being organised hierarchically, with the simpler processes such as reflex situated at the lowest or subconscious level, with progressively more complex processes such as sensation, perception, abilities of association and relation defining the progressively higher (conscious) levels of function (Verster, 1987). It will become apparent that Spencer's ideas which he postulated well over a century ago are remarkably close to the ideas we have about human intelligence even today.

Sir Francis Galton, a half-cousin of Charles Darwin, who has come to be known as the 'father of mental tests' was the forerunner of the scientific approach to the study of human intelligence (Kail & Pellegrino, 1985). Galton was the first to develop the notion

of individual differences which has become the central dictum of the psychometric tradition today and was also the first scientist to postulate the existence of a "general mental ability" in humans (Walsh & Betz, 1985).

Galton's work was strongly influenced by his cousin's theory of evolution. He sought to demonstrate the hereditary basis of human genius, proposing the notion of a genetic base for human ability, which he attempted to do through a study of men of influence by venturing to prove that the extraordinary abilities of many men are passed on in some way to their offspring. Galton also held the view that differential environmental pressures operating over time on different segments of the population resulted in the differential intellectual development of racial groups (Verster, 1987). Put simply, Galton introduced the idea that biological inheritance is key in determining intellectual ability. In addition, he proposed that the intervention of environmental factors has an important impact on the development of this ability.

These notions were the nucleus of the nature / nurture debate which has become such a central issue in socio-psychological thought, and have a direct bearing on the argument against the use of uniform psychometric tests for different populations in South Africa, a topic which will be entered into later.

Galton's belief in the biological determination of all mental abilities and his techniques for measuring human ability also led to the development of important statistical concepts which have had a profound effect on the development of subsequent psychometric methodology (Verster, 1987).

For example, Galton predicted that mental phenomena would follow the same law of inheritance as physical phenomena such as eye colour, height, etc. He proposed that just as physical attributes such as height were known to be normally distributed in the general population according to a symmetrical bell-shaped curve, so intelligence too if measured across the population would conform to the shape of a normal distribution. This gave rise to the first normative psychometric scales which describe individual differences in intelligence and other abilities in terms of units of standard deviation from the mean (Verster, 1987). Galton thereby greatly extended the range of statistical procedures that could be applied to the analysis of data.

J.M. Cattell (1860-1944), the first to coin the term "mental tests", joined Galton and together they developed numerous laboratory devices to measure sensory capacities which they thought were the components of intelligence (e.g. visual and hearing acuity, reaction time, memory span). As a result of this focus on simple sensory-motor components of behaviour, a different viewpoint began to arise among psychologists that it was complex and not simple processes that were the key to predict "intelligent behaviour" (Hughes, 1989).

Galton's work precipitated significant advances in the field, most notably by the Frenchman, Alfred Binet, who in 1905 devised the first tests of higher cognitive processes (Cronbach, 1949). Like those before him, Binet held the view that the mental processes underlying intelligence are multifaceted and hierarchically organised, but emphasised the importance of environmental opportunities for learning in the development of intelligence. The belief that changes in cognitive capability correspond

to chronological growth and that intellectual development in childhood involved the adaptation to environmental demands by mastering each form of cognitive functioning, began to gather momentum. It can also be seen at this point that there began to develop an awareness of the critical role that social, educational, and economic environment plays in the development of cognitive capability.

The work of Spencer, Galton, and Binet encapsulated efforts to use tests to determine levels of intellectual development. Another thrust of the testing movement was concerned with the use of tests to generate theories of intelligence and individual differences. It was in this vein that the next important advance in the development of cognitive psychology came, with Spearman's development of the statistical method of factor analysis (Anastasi, 1982).

Factor analysis is concerned with the analysis of relationships among tests through correlation. If one had to examine, for example, the abilities of a pool of several athletes from a range of different activities such as sprinting, marathon, long jump, discus, etc. one could say that every athlete possessed general athletic ability. However some would be highly developed in speed and power events while others would excel in endurance events. By putting all the athletes through each type of event and correlating the performance of each one on a matrix of abilities, it would be possible to isolate their specific abilities. An ability is isolated when a number of results of performances correlate strongly with each other. This is essentially what factor analysis is concerned with.

In 1904 Spearman used a statistical process which had been discovered by Pearson a few years earlier called product-moment correlation to uncover the major areas (or 'factors') of human mental capability (Verster, 1987). Pearson's discovery of correlation made it possible to predict how accurately one variable could predict the behaviour of another. Spearman deduced that correlation could be interpreted as representing some common factor of causation amongst two variables. By analysing the inter-correlation of test results, Spearman was able to estimate the amount of common variation between two variables and extract the common ability or 'factor' that existed in both of them.

As a result of his discoveries, Spearman proposed a two-factor theory of intelligence. He suggested that there exists an all-embracing general intelligence factor which he called 'g-factor'. He considered the 'g-factor' to enter into all tests of mental performance, but is most purely measured by tests demanding the use of higher functions such as logic, judgement, and reasoning.

The other factor, which he identified as 's-factor', reflected the specific factor which was peculiar to each test. The two-factor theory therefore postulated that in cognitive functioning there exists a general factor of intelligence ('g-factor') which is complemented by many specific abilities, each being unique and different from the other, as well as being separate from 'g-factor' (Aiken, 1982).

Each task that a person performs requires a general ability as well as the specific ability to be able too manage that specific task. For example, some common ability would be required for success in science and mathematics. These subjects require competencies

which may be termed numerical ability. But whatever this ability may be termed, it would not be required to the same extent for an individual to be successful in English or history. Success in this subject would require a different type of ability, which might be termed a combination of verbal ability and memory.

However, a student may be competent in all three subjects, which would suggest a good general ability as well as high levels of specific abilities.

The notion of a hierarchical organisation of the principle components of mental ability was extended by Cyril Burt (1949). Through the application of factor analysis, he put forward evidence for the existence of at least eighteen factors of intelligence located on successive levels below 'g' (Verster, 1987). In the same vein, Thurstone, in 1931, administered 56 tests to 240 students at the University of Chicago, and factor analysed the matrix of correlations between the test results, revealing nine separate factors which he labelled "primary mental abilities" (Verster, 1987:16). These were identified as spatial ability, perceptual speed, a number facility, verbal reasoning ability, rote memory, word fluency, inductive reasoning ability, and the remaining two factors he tentatively identified as deductive reasoning ability, and a factor which he called the 'restriction' factor (Sternberg, 1985).

A major hierarchical theory of intellect which represented an extension of Burt and Thurstone's ideas was that of Vernon (1965). It is necessary to unpack the various factors which he identified in order to put into perspective the constructs which modern tests are designed to measure.

This theory was also based upon a factor-analytic approach which starts at the top of the hierarchy with Spearman's 'g-factor' as the most general factor of intelligence. The model follows a top-down approach with each level having a more and more distinct and isolated (as well as limited) ability than the preceding level (Kail & Pellegrino,1985).

Two major group factors of intelligence fall under the "general cognitive factor 'g': a verbal-educational ability (designated v:ed), and a practical-mechanical-spatial group factor, designated k:m" (Aiken,1982;194). These two major group factors represent the broad categories that are thought to comprise human cognitive ability.

The major group factors then divide into minor group factors, each denoting a specific ability. The v:ed factor can be further decomposed into verbal and numerical ability. Similarly, the k:m factor is broken down into manual ability, mechanical information, and spatial ability (Guilford,1978). In addition, there are certain cross-links such as mathematical ability which are influenced by both spatial and numerical ability. This decomposition of factors can be continued to derive even more levels in the hierarchy, the most important of which will now be described.

2.1.2 Perceptual Speed Factor

This factor concerns the ability to see assimilate visual patterns and detect the relationships between them, identifying similarities and differences quickly and accurately (Nunally, 1970). An example might be a row of designs with the testee having to spot the error. A person with high perceptual speed would identify the error in a very short time.

2.1.3 Verbal Comprehension Factor

This factor entails the ability to understand ideas expressed in words and is the principle factor employed when performing tasks such as reading comprehension, and solving verbal analogies and disarranged sentences (Anastasi,1982).

Nunally (1971:240) explains that "verbal comprehension represents most of what we refer to as 'reading skill' ". Verbal comprehension would usually be assessed through tests which measure knowledge of synonyms or antonyms, or similar. A typical example of a test item for measuring this might be: "Find the word which has the same meaning as ANCIENT ". The testee will then have top choose the correct answer from a given list of:

A. Dusty B. Age C. Old D. Historical

2.1.4 Number Facility Factor

This factor involves the ability to solve arithmetic problems rapidly and accurately. It is typically measured by arithmetical word problems in which there is some emphasis on both computation and reasoning, but relatively little emphasis on extenet of prior learning (Sternberg, 1985).

2.1.5 Induction Factor

This could be described as general reasoning ability as inductive reasoning tasks are viewed as measures of a more general ability. This factor is measured in tests that require the determination of some rule or relationship in a pattern of figures, series of

numbers, or letters, and requires the ability to apply logic in order to solve problems where the inference of a principle or concept is demanded (Kail & Pellegrino,1985).

2.1.6 Deduction Factor

Nunally (1970) explains that the deduction factor is concerned with the drawing of conclusions, as in logical syllogisms (a form of reasoning in which from two propositions a third is deduced). An example of the type of item used to measure this factor may be:

"A is taller than B, who is taller than C. Who is the shortest?"

2.1.7 Spatial Factor

This is one of the most analysed and discussed factors in the area of mechanical aptitude. Wiseman (1967:207) asserts that "mechanical ability involves general intelligence, the spatial factor, and to a smaller extent the manual and numerical factors, and possibly some such factor as perceptual and motor speed."

Spatial ability is thought to represent the ability to visualise how parts of objects fit together and what they would look like if rotated in space. The spatial factor is assessed by tasks that require the individual to judge whether two items rotated in space are the same or different.

Spatial aptitude falls under the spatial-practical-mechanical construct (k:m) in Vernon's hierarchy. It can be divided further into three factors (Kail & Pellegrino,1985):

(i)Spatial orientation appears to involve the ability to imagine how a stimulus would appear from another perspective.

(ii)Spatial relations involves the ability to rapidly and accurately in mental rotation processes in order to decide if two stimuli are identical.

(iii)Spatial visualisation is detected by tests in which the stimuli are more complex than those on tests of spatial relations and in which the speed with which individuals respond is less emphasised than the number of problems correctly resolved. Such tasks would typically involve the mental folding or unfolding of flat patterns. For example, a flat piece of paper is folded in a described manner, and the testee must choose which one of the choices represented is the same as the original. This would require the testee to mentally fold the piece of paper in his or her mind.

2.2 Recent theories

More recent thinking has diverged from the earlier fairly minimalistic idea of intelligence consisting of specific, discrete factors which can bemeasured numerically. More current ideas advocate that intelligence should be viewed as a dynamic, evolving entity.

R.B. Cattell (1963) proposed his theory of "fluid" and "crystallized" intelligence which has much in common with Vernon's two major group factors k:m and v:ed. He considered "fluid intelligence" to be a "relation perceiving capacity and is dependant on the neural-physiological intactness and efficiency of the brain" (Maloney & Ward, 1976:182). Fluid intelligence is measured by tests which require an ability to see relationships such as number and letter series and verbal analogies.

"Crystallized" intelligence is considered to be the result of fluid intelligence and the environment interacting, and is therefore sensitive to cultural, educational and environmental influences (Maloney & Ward, 1976). It is measured by tests using content which draws on previously acquired knowledge such as vocabulary and arithmetic (Walsh & Betz, 1985).

2.2.1 Developmental theories

Piaget formulated a theory of intellectual development based on a slow evolving of intellect over the years. His theory of intellectual development comprises of three stages: a sensorimotor stage (birth to beginning of language); a concrete operations stage (roughly two to eleven years old, where language and thought becomes more advanced); and a formal operations stage (from eleven years onwards, where a child develops a capacity for reasoning and self-representation) (Hughes, 1989).

2.2.2 Learning theories

Learning theory diverged completely with the idea of intelligence comprising of traits. Essentially, what this theory postulates is that a theory of intelligence is a theory of learning. The nature of intelligence is thought to lie in the learning process itself (Maloney & Ward, 1976). In a broad sense, intelligence is viewed as an end function of the learning process functioning at its optimal level (Hughes, 1989).

Summary

This chapter has attempted to trace the development of the theories underpinning psychological measurement from the earliest theories to more contemporary ideas. It is

clear that the thinking surrounding psychological measurement was reductionist in approach, identifying specific abilities and measuring them in isolation from each other. It will be attempted to show that many of the tests that are currently used today in industry, including those employed in the selection battery in this study, ascribe to the same principles of assessing the individual's abilities and intelligence according to rigid, discrete factors.

There have been attempts in recent years to develop psychometric tests which do not attempt to measure specific aptitudes and abilities, but are rather oriented to measuring the ability of the individual to learn and rate of the individual's ability to absorb and apply new concepts. There appears to be movement towards those tests which measure potential or innate ability and not previously acquired static knowledge. The rationale behind this development is that tests of learning potential, or what Cattell would have called 'fluid' intelligence, are a more objective and "culture-fair" assessment of the individual's capacity to perform in future after exposure to training, and as a result will not attract the scrutiny and suspicion of those tests which measure static abilities, or 'crystallized' intelligence which may have been constructed due to environmental, cultural, educational, or social influences.

A following chapter will examine the psychometric tests and the trade test themselves in the light of what has been discussed in the current chapter.

CHAPTER THREE

3.0 Testing considerations in South Africa

Any investigation regarding psychometric testing and the use of assessment instruments for personnel selection within the current environment of South Africa today must consider the context of testing in this country.

The frailties inherent in traditional psychometric testing instruments as a result of the philosophies giving rise to test design have been expounded upon, but the implications thereof must be mentioned in order to understand their use in the South African context. Once the contextual considerations have been discussed, the test results will be reviewed in this light.

Reference to psychometric testing is made in the new Employment Equity Act tabled recently, and the message carried therein is clear. Chapter two (point 8) reads:

"Psychometric testing of an employee is prohibited unless the test being used -

(a) has been scientifically validated as providing reliable results which are appropriate for

the intended purpose.

(b) can be applied fairly to employees irrespective of their culture; and

(c) is not biased against people from designated groups."

The inference from the above is that the unfair application of assessment procedures could constitute an unfair labour practice in terms of the Act . One could also draw the inference that by not affording candidates equal assessment through equal assessment, one would in principle be committing an unfair labour practice (Erasmus, 1995)

Guidelines which assessment procedures should meet have been suggested by at least one of the trade unions. It could be expected that labour disputes involving the fairness or unfairness of assessment instruments will largely be based on these guidelines (Erasmus, 1995).

The National Union of Metal Workers of South Africa submitted a proposal to employers in the Iron and Steel, Motor and Automobile Manufacturing Industries to end unfair employment practices which further emphasises the approach that government and unions are taking. Extracts from the document state that,

" 4.1.2 The tests shall be relevant and properly analysed for the job for which the applicants are being considered

4.1.4 The test should be valid for the purposes of selection. The validity of the tests with different groups should be investigated empirically

4.1.5 Apart from conventional validation, job-analysis-validation should be done and this procedure should also be used to set realistic cut-offs

4.1.6 The bias in tests shall be assessed. Predictive bias studies should be undertaken

4.1.7 If the applicant is tested in a language that is different from his / her home

language, a non-verbal assessment should be included. A correction factor must also be built into the interpretation of the test score, for example additional points or additional time allocation

4.1.9 Tests of learning potential, as well as job specific tests such as work sample tests or trainability tests should be incorporated

4.1.11 Separate norms for different groups should be used until such time as the bias

in tests is accounted for, or different groups of applicants have more equitable past experience

4.1.15 In order to compare test scores across applicants it is necessary to consider the developmental opportunities to which they have been exposed. As race groups in this country have lived through very different and inequitable learning environments, it is debatable whether test scores are comparable for all applicants. Currently there is not a great deal of research data available in South Africa to ascertain the comparability of test scores among race groups, thus where tests continue to be used, employers should undertake to investigate this."

(NUMSA,1992)

There is obviously a general concern about the problems with and applicability of Westernised tests within the South African context. Taylor and Radford (1986) discussed how vital it is to address unfair labour practices, particularly in the area of psychometric testing. They argued that psychometric testing in South Africa is a means of entrenching gross unfair labour practices as conventional tests have been designed

along western standards, standardised on one population group only, and are being applied to non-western cultures. They concur that tests of the same kind administered to different ethnic groups led to differences in mean scores.

Taylor and Radford (1986) also concluded that a concerted effort has to be made in order to gain more information about the comparability of psychometric tests for different ethnic groups, and for policy makers to be aware of the influence of psychometric testing on selection fairness.

Differential performance between racial groups on psychometric tests and the resulting debates around "unfair discrimination" and "selection fairness" has become the most controversial and heatedly debated issue in the field of testing (Cole, 1981). There have been well supported cumulative research findings substantiating the existence of group differences (Spies-Wood, 1998), where blacks have tended to score one to one-and-a-half standard deviations below whites on cognitive tests (Dreger & Miller, 1960, 1968). Such findings have generated a considerable amount of controversy and dispute over the issue of cross-cultural assessment (Barrett & Bass, 1975). The issues confronted when comparing the meaning of test scores must be raised because of the relevance of the debate to the study in question, since all four psychometric tests used in the selection battery were administered equally for all race groups. The complexities of the argument will not be entered into because of the depth and extent of the differing positions.

Retief (1988) cautions that when a test is used in another culture, problems of comparability, equivalence, translation and internal bias arise. Poortinga (1983) explains

the differences in test performance between groups by his notion of "cultural distance". He defines this as the behavioural repertoire or "set of response patterns at the disposal of a person belonging to a certain race group. When the overlap in repertoire between two groups is large, the cultural distance between them is small, with the result that tests may demonstrate negligible unfairness" (Wheeler, 1993:27). The situation in the United States is fairly indicate of this since the process of acculturation and socio-economic upliftment of disadvantaged minority groups is far advanced (Ibid).

Anastasi (1982) argues that all behaviour is affected by culture and that cultural influences will always be reflected in test performances. Differences in test scores have therefore been attributed to the inclusion on "culturally loaded" items (Arvey & Faley, 1988). "Culture-fair" tests aim to eliminate the unwanted effects associated with cultural differences by ruling out such parameters along which cultures vary such as language, reading and speed (Anastasi, 1988). A factor common to culture-fair tests is a decreased emphasis on verbal items and instructions, and a greater emphasis on processes such as visualisation, spatial and abstract reasoning (Ibid).

There are a number of concerns regarding the use of the four tests used in the selection battery with regards to their applicability across all racial groups. The Mechanical Comprehension test and the Intermediate battery Mental Alertness test both have a large verbal component and draw on knowledge that would have needed to have been acquired through formal education and socialisation. These tests could be said to be a measure of "crystallized intelligence" since they measure static, previously-developed abilities. The applicability of these tests to all groups must be questioned. Conversely, the High Level

Figure Classification test and the Blox test are essentially non-verbal tests of spatial and abstract reasoning, or "fluid intelligence" and therefore would appear to be more culturally fair than the other two tests.

The element of language is the obvious difference between the two "types" of tests used in the battery, and can be seen as the greatest factor inhibiting the cultural fairness of the test battery.

Campbell (1995) investigated the effect of language on psychometric test performance of a Black sample group. The results are interesting and also applicable to the current study because test instructions are given in one language only (English) and in two of the four tests used - Mental Alertness and Mechanical Comprehension - English is the test medium.

Campbell (1995) argues that equivalence in translation is of paramount importance. He states that it is not entirely possible to translate a question into another language and have the same meaning. One could argue further that in the case of speed tests (i.e. having a limited time period in which to finish), and all the four tests under consideration fall into this category, time taken in translating a question item into another language must have a strong bearing on test performance. When one considers that the effect of such translation may results in perhaps three questions not being answered out of thirty items because of time constraints (this is a hypothetical estimation), the liability becomes obvious.

Campbell (1995) undertook to demonstrate that a fundamental reason why Zulu speakers perform less well on psychometric tests (on average) than their English speaking counterparts is not due to lack of intellect, but rather a matter of misinterpretation of the English based questions.

After administering a general mental ability test to sixteen university students, the results were classified into ten groups of categories / reasons. The fundamental questions that Campbell was attempting to address were: "Do the testees really understand what the question is asking of them?" and secondly, "Do they interpret the questions in the same way as a westernised individual would?"

Conclusions drawn were that certain questions in the test were answered wrong by over half of the students. This indicated that the answers were incorrect not due to negligence, but rather to "some other intrinsic factor deeply rooted with the way then questions came across and were interpreted by the respondents." (Campbell, 1995:101).

More interesting is the fact that a large number of the test questions items were based on verbal reasoning systems that required the respondents to be familiar with vocabulary, its meaning, and the context in which it is used (Ibid). This is especially significant in the case of the tests under investigation, two of which have a large verbal component.

Some explanations are offered in terms of some of the factors that could have led to misinterpretation. The first concerns the use of pronouns instead of nouns. The use of pronouns "may have lead to unclear references due to vague non pronoun links" (Lonner

& Berry, 1986:145 in Campbell, 1995:102) as in the case of the first test item: "A party consists of a man, his wife, his three sons and their wives and their three children in each of the son's families. How many were there in the party?" A confusing sentence to unravel in one's own language, the difficulties are surely exacerbated when posed with such a question in a second or third language.

The use of the subjunctive was another factor leading to misinterpretation, that is the use of "could", "would", "should" etc. (Lonner & Berry, 1986 in Campbell,1995:102) explain that there is often no readily available term in other languages for the various forms of the English subjunctive.

A third area that lead to misinterpretation was found to be questions that required the use of the alphabet to decipher a word written in code. It was proposed that the fact that the alphabet is not present in the question may lead to problems in solving the question. It may also be valid to state that historically, certain groups in this country have been taught from the outset of primary school to become familiar with the order and pronunciation of letters in the alphabet, perhaps more so than other groups.

In the light of the context surrounding psychometric testing currently in South Africa, there are a number of fundamental problems in applying tests that have been developed for and standardised on one population group being applied to other groups which have had different environmental, educational and social experiences.

Summary

Some of the potential problems associated with the four tests in this investigation have been outlined in this chapter. The social and legal environment in South Africa has become intolerant of selection instruments which cannot demonstrate that they are free from all cultural influences. Differential performance between groups on tests has been explained in terms of influencing factors such as language, and in terms of this, it must be recognised that two tests used in the battery (Mental Alertness and Mechanical Comprehension) make use of verbal test items. The High Level Figure Classification and the Blox tests utilise figural and diagrammatic test items and therefore, and therefore would appear to be more culturally fair.

An investigation conducted by Holburn (1992), dealing specifically with test bias in the four psychometric instruments used by the Durban City Council will be examined later, and the findings thereof may then be viewed against what has been discussed in this chapter.

The following chapter will discuss the tests used in the selection battery and their properties, and the trade test will then be described.

CHAPTER FOUR

4.0 The assessment criteria

Personnel selection procedures are used to "predict future performance or other behaviour," and evidence for this criterion-related validity typically consists of a demonstration of a "useful relationship between the selection procedure (the predictor) and one or more measures of job relevant behaviour (the criterion)" (The Guidelines, 1992:11)

The Guidelines (1992:44) also state, "whenever any selection procedure is used, it is used at least with the implicit assumption that some aspect of behaviour on the job (including performance in training) can be predicted from numerical scores on that selection procedure. The essential principle in the evaluation of any selection procedure is that evidence be accumulated to support an inference of job relatedness."

The predictors and criterion used in the selection and evaluation of the apprentices will be examined. The key concerns will be firstly to determine the rationale or reasoning behind the use of the selection instruments in light of the use to which they are being put. The issue here is essentially one of 'face' validity - do the tests in question seem to be a logical methodology of assessing a candidate for the training course? It is obvious that a predictor is more likely to show validity if there is a good reason to suppose that a relationship exists between a predictor chosen and the behaviour it is designed to predict (Society for Industrial Psychology, 1992).

In addition, it should be asked whether there appears to be an 'overlap' in content between the predictor and the criterion i.e. are both instruments measuring similar competencies or attributes? It is naturally desirable that the predictor is relevant for the criterion, that is, is it assessing the same domain of competencies or attributes as does the criterion?

The reasons for reviewing the psychometric tests in this light is because ultimately the predictive validity of an assessment instrument is dependant on the test content and the similarity between the test content and the criterion content. Importantly though, the Guidelines (1992:29) state that, "the acceptability of the match between selection procedure and job content domain is a matter of professional judgement".

It is however important to remember here that the content of the criterion is not under scrutiny, nor is its appropriateness, thoroughness, or even fairness. The criterion is the fixed variable in the equation, the predictor is not. The electrical trade test has been and will likely remain the established means for assessing the competence of an apprentice after completing his or her training. In addition, the selection predictor may be changeable, the criterion measure however, will not. It is therefore the benchmark against which any selection instrument must be matched, not vice versa.

4.1 The predictors

In 1994 the psychometric test battery used in the selection of artisan apprentices comprised of the following tests:

* Intermediate Battery Mental Alertness Test
* Blox Test
* High Level Figure Classification Test
* Mechanical Comprehension Test

Each of these tests will be described in detail below.

4.1.1 Intermediate Battery Mental Alertness Test

This is a sub-test of the Intermediate Battery (B/77) which includes other tests in addition to the Mental Alertness. There also exists a separate High Level Battery, which would be aimed at a cognitively higher functioning group. The Intermediate Battery Mental Alertness alone was administered.

This test was designed as a measure of general reasoning ability in the following three areas: quantitative-analytical ability, verbal ability, and clerical speed and accuracy, thereby giving a generic impression of an individual's abilities and aptitudes (Administrator's Manual B/77:1973).

The test was designed to reveal verbal and numerical reasoning ability and is measured by verbal analogies, completion of number and letter sequences, decoding problems based on the alphabet, and the identification of similarities between words and ideas.

Thirty test items must be answered in 30 minutes. The answers are in multiple choice format with five possible choices.

4.1.2 Blox Test

Formerly known as the Perceptual Battery, the Blox test, according to Holburn (1992), is a measure of the individual's ability to recognise spatial arrangements from different angles, an ability considered to be a subset of spatial ability (Halstead & du Toit, 1983).

Spatial relation and orientation are some of the factors which make up spatial ability, and are described as "the ability to comprehend the nature of arrangements within a visual stimulus pattern primarily with respect to . . . one's own body or frame of reference" (Michael, Guilford, Fruchter & Zimmerman, 1957:57). The Blox Test Administrator's Manual (NIPR:1983) states that spatial visualisation, itself a factor of spatial ability, involves the ability to mentally manipulate (i.e. rotate, invert, or twist) one or more parts of a visual stimulus pattern and to recognise the changed appearance of the object. The manual states further that the Blox test, which measures the ability to recognise spatial arrangements from different angles without the benefit of a physical shift of the body, is a measure of spatial relations, visualisation, and orientation.

According to Michael et al (1957), this test was designed to measure the ability to perform mental manipulations of three dimensional visual images which appear as cubic clusters,, and to understand the nature of the arrangements of the elements contained within this visual stimulus pattern, and is therefore essentially a measure of spatial relations.

The Blox test consists of forty five items and six practice items which have to be answered in 30 minutes. Each page is divided in two by a line. Above the line five

drawings are presented, each figure containing combinations of between two and six blocks (or cubes) linked together in a 'cluster'. In other words each figure represents a number of three dimensional cubes stuck together. Below the line are nine sets of figures of the same kind. The figures below the line provide the possible responses, and the figures above the line are the stimuli for each test question. Naturally, five of the figures below the line are identical to the five stimuli above the line, but are represented so as to be seen from a different angle or perspective.

The task of the testee is to recognise each figure above the line within the nine figures below the line.

The parts that comprise each figure maintain their relationship to one another as the whole figure is mentally rotated in space. So, as each cluster of cubes is rotated in the testee's mind, the cube itself retains the same shape. In order to identify which of them nine figures below the line is the exact replica of the item he / she is looking at above the line, the testee has to imagine (by looking at the two dimensional figure) what it would look like from another angle. To do this the testee needs to imagine the figure as a three dimensional object and mentally move it about, looking at the figure from different angles.

4.1.3 Mechanical Comprehension Test

This test was designed to measure mechanical comprehension - the ability to apply knowledge of the laws, principles, operations and effects of physics appropriately, and

"should therefore be useful in the selection of personnel in many technical fields" (Administrator's Manual, 1968:1).

Items are drawn from the field of applied mechanics, general physics (electricity, fluid mechanics, optics, thermodynamics, acoustics and magnetism), spatial relations and engineering practice (Ibid.). Items in the test would comprise of a drawing involving a representation of a pulley, gears, object (for example a magnet) or structure (e.g. a bridge). A test question would typically ask that if an action were to be performed, what effect it would have on an object (e.g. heat or cold applied, cog rotated, etc.). Success in the test depends to a large extent on the testee's knowledge of the laws of physics which would be acquired at secondary school level.

The test contains forty-two multiple choice items, each consisting of an illustration followed by a brief clarifying comment where appropriate, and three possible answers. The time limit is 35 minutes.

4.1.4 High Level Figure Classification Test (HL-FCT)

The HL-FCT was developed out of a need for a non-verbal test of general intelligence suitable for all cultural groups (Holburn, 1992). This test is a non-verbal assessment instrument designed to measure abstract and conceptual reasoning ability uncontaminated by verbal skills. It is contended by Holburn (1992) that the HL-FCT provides a good index of an individual's general intellectual ability. It was originally intended for use in selecting people for positions which demanded a moderate level of abstract conceptual functioning i.e. work that involved more than mere routine activities.

There are twenty four items which are to be answered in 30 minutes. Each test item comprises of six drawings labelled from A to F which must be classified into two groups of three so that members of each group share a common characteristic or are similar in some way. The testee indicates his or her answer by recording the three letters of each group on an answer sheet eg. AFE BCD. Each test item was designed in such a way that it is always possible to form two groups of diagrams out of the set of six.

Each test item is based on a different concept on which the classification rule is based such as the direction in which the figures face, the size or shading of the figures etc. The testee is required to identify the new concept which underlies the test item and group the figures according to this rule.

4.2 The criterion

The trade test which apprentice electricians undergo at the end of their training consists of a range of sixteen possible tests or tasks which examinees are required to perform. The examiner therefore has an option of sixteen possible questions to ask the candidate, six of which are chosen, including a compulsory test of manual dexterity. The intention of the trade test is to measure as accurately as possible within a relatively short space of time the broad range of knowledge an apprentice has acquired over approximately two years of training.

The analogy of the matriculation examinations clearly identifies the problems involved. The question could be posed, for example, whether it is possible, or indeed fair, to

attempt to assess a matriculant's understanding of the year's history syllabus through a single three-hour examination.

It must be acknowledged firstly that it would be impossible to test a candidate's knowledge of each area of the syllabus, nor would it be possible to do justice to the sections that are concentrated on by the examiner. It would also have to be accepted that the candidate who concentrated a disproportionate amount of energy on only a few areas of the syllabus may be fortunate enough to be examined on these same areas.

Conversely, the candidate who has a fair general understanding of the entire syllabus may be found wanting because the questions asked in the examination were focused on only a few areas, and the examinee therefore was not able answer those questions in enough detail.

These are the dilemmas which the trade test similarly attempts to overcome. However, taking the matriculation example again, it would be true to state that in general, those matriculants who obtain the highest symbols have been the top students throughout the year, and possess a good understanding of those subjects relative to others; and those who have struggled through the year usually find the examinations difficult. In summary, the final matric symbols are by and large a relatively fair summation of each candidate's academic abilities.

The results of the electrical trade test support this generalisation. According to an experienced instructor at Durban Electricity's training school, it can be reasonably

predicted how well an apprentice should do on the trade test on the basis of his work throughout the year, and this view was widely held by other instructors. The instructor confirmed his faith in the trade test as a fair and thorough assessment of each apprentice's overall ability due to the fact that there are rarely any "surprises" in the results - in general those apprentices who show the greatest flair achieve the highest results, and those who struggle tend to find the trade test very challenging.

4.3 The trade test

It is not possible to gain access to the actual testing situation in order to gain first-hand experience of what the trade test entails. However, Durban Electricity has set up a comprehensive simulation centre comprising of all the elements of the trade test and operations which need to be carried out. The apprentices are encouraged to undergo "pilot runs" at the training centre under examination conditions, using past trade test questions, with the same time limits as would be used in the actual examination. This is done in order to prepare the apprentices for the actual test.

Each question in the trade test requires the candidate to perform a task at a work station or cubicle. The work station comprises of a wall mounting housing the electrical device or panel which forms the basis of the test question. The tools and instruments required to answer the question are provided.

It was possible to gain permission from the Department of Electricity to observe the apprentices engaged in realistic simulations of the trade test. Each apprentice was observed answering the questions on the trade test and carrying out the tasks required.

The observation was carried out with a qualified trainer who would advise the author on exactly what the question was asking, and commented on how the apprentice was responding to the question. Of special interest was observing the difficulties experienced by the candidates and the techniques used to solve these problems. The apprentice was evaluated at the end of the test. Each apprentice was observed doing the pilot run twice.

The reason for observing the pilot runs was to gain an understanding of the domain of competencies or attributes which apprentices would need to possess in order to succeed in their training. The intention was to compare the competencies required for successful completion of the trade test against the domains measured by the predictor tests in order to assess whether there were obvious similarities in the competencies or aptitudes measured by both the assessment criteria. In other words, this entailed a rudimentary analysis of the "face validity" of the psychometric tests.

4.4 Identification of competencies

The most significant dimension appeared to be the ability to apply logical, systematic reasoning to a problem. After reading the test question, the candidate would embark on a process of diagnosis which usually entailed establishing a cause and effect relationship between two elements of the problem, i.e. "if this is not working then the problem must lie here". The ability to trace the source of a problem through deductive reasoning, or *fault-finding ability* was identified as the critical competency. It could also be said that this would be the activity in which a qualified electrician would be mostly engaged in the field.

The ability to read and clearly understand instructions, although not a "key performance indicator" played a significant role in how quickly a candidate moved through a problem. It was apparent that much time was often wasted because a candidate was obviously struggling to understand what the question was asking, and even more time spent reversing a process and restarting once it was realised that the question was being approached incorrectly.

In order to perform a number of the tasks, a candidate would be presented with a drawing of a circuit or a panel. This diagram would be a representation of something that would need to be constructed by the candidate, or it would be a representation of the panel or circuit which was present before him or her. The candidate would be required to use the schematic diagram as a reference point to perform some task with a physical tool or instrument. The ability to interpret a two-dimensional sketch diagram and apply it to a three-dimensional circuit on the wall mounting was therefore seen as important.

Mathematical ability was observed to be employed on a number of occasions during tasks as candidates made calculations and applied formulae in order to solve pure mathematical problems, especially in one test specifically consisting of resistance circuit calculations.

Practical hand-eye co-ordination was deemed to be important as was manual dexterity. Candidates would frequently be required to work with small objects such as bolts, strip and cut small wires, and insert wires into the correct holes on a circuit board. Much emphasis is placed on manual dexterity, to the extent that a compulsory test called

'Installation Work' which assesses this ability has to be completed by every candidate. This test requires the candidate to bend, hammer, and twist a variety of pipes and tubes using heavy tools and to install these in a specified manner. The object of the test is to be able to assess how quickly and dextrously an individual can manipulate some fairly resistant materials such as metal piping. This test requires a degree of physical effort even though some rather powerful and bulky tools are being used.

Using the terminology introduced in chapter two when discussing the hierarchy of abilities, the domain of competencies could be described in these terms:

4.4.1 Verbal ability

There is a verbal comprehension component within the trade test. Examinees are required to read and understand written instructions and apply the instructions to a task.

4.4.2 Number facility

There is a number facility component since examinees are required to perform mathematical calculations.

4.4.3 Induction factor

An induction factor is present. Examinees have to apply logic in order to solve problems where the inference of a principle or concept is demanded.

4.4.4 Deduction factor

A deduction factor is present since an examinee must be able to draw conclusions through applying a form of reasoning by which, from two propositions a third is deduced, e.g. "if the instrument is giving me such a reading and I have already adjusted this component, then the logical deduction is that fault must lie here ."

4.4.5 Spatial ability

A spatial ability exists. Examinees must be able to assimilate a two dimensional diagrammatic figure and apply it to a three dimensional construction. This would require the mental manipulation of the drawing.

It is of interest to compare the domains of aptitudes identified here to the abilties identified by Biesheuvel (1949) that an ideal apprentice would possess, namely:

(a) High overall intelligence

(b) Scholastic aptitude especially in mathematics and science

(c) Spatial ability

(d) Mechanical insight

Steyn and Latti (1974) identified the following parameters as important selection criteria for first year engineering students at the University of Pretoria:

(a) General mental alertness

(b) Deductive reasoning

(c) Inductive reasoning

(d) Numerical ability

(e) Spatial ability

(f) Perceptual ability

Many of the aptitudes identified as necessary for achievement by the above authors (especially Steyn and Latti, 1974) have been identified in the analysis of the trade test.

Consider the domain of competencies which the trade test isolates for assessment. These criteria should be the reference points against which any predictor instrument should be measured in terms of suitability. Fundamentally, the selection instrument should assess the same competencies as the criterion measure.

The Guidelines (1992:44) state that, "When any selection procedure is used, it is used at least with the implicit assumption that some important aspect of behaviour on the job can be predicted from numerical scores on that selection procedure. The essential principle in the evaluation of any selection procedure is that evidence be accumulated to support an inference of job-relatedness." The predictors must be assessed in the light of this statement.

The Mental Alertness test is a measure of general reasoning ability, but isolates the following factors specifically: verbal ability, quantitative-analytical ability, and numerical ability. This test would seem to have high face validity when predicting for achievement on the trade test since the ability to read and understand, to calculate, and to apply quantitative-analytical reasoning skills are required in the trade test. However, the

high verbal component of the test questions the appropriateness of the test for testees for whom English is not a first language.

The Blox test is a measure of spatial ability. This aptitude was identified through the analysis of the trade test and also by Biesheuvel (1949) and Steyn and Latti (1974) as important factors for apprentices and engineering students to possess. The trade test requires the candidate to interpret two-dimensional drawings and apply them to three - dimensional structures, by mentally manipulating (rotating or inverting) the visual stimulus. The Blox test could be said to have high face validity for the purpose of indicating success on the trade test.

The Mechanical Comprehension test is essentially a measure of the testee's knowledge of and ability to apply the laws of physics to certain situations. Ability to perform on this test is dependant on schooling and socialisation. None of the aptitudes identified in the trade test are isolated for measurement by this test. This test would appear to have little or no face validity for the purpose of indicating success on the trade test.

The High Level Figure Classification Test is a measure of abstract and conceptual reasoning ability. Inductive and deductive reasoning are requirements for the trade test, and although these specific forms of reasoning are not measured by the HL-FCT, one might expect that it would have some face validity when used to select for the trade test.

Summary

The content of the predictor tests and trade test has been described and analysed in some detail. An attempt was made to identify the competencies which each assessment

criterion measures in order to see if there is any overlap. If this occurred, it was deduced that on a non-scientific basis, the predictor appeared to be have a relationship with the criterion. Three of the psychometric tests were shown to possess face validity or appropriateness for the purpose of indicating success on the trade test.

The next chapter will provide and overview of other research which is related to the present investigation. The validity of the predictor tests in various roles is examined, and the shortcomings of the test in terms of bias is discussed.

CHAPTER FIVE

5.0 Review of related research

Holburn (1989) investigated currently used apprentice selection methods with a view to developing or adapting appropriate techniques. Six hundred and forty-four organisations which employ apprentices were assessed. The greatest number of apprentices emplyed in industry were found to be in the electrical (heavy current) category (25%), the next two largest categories being maintenance mechanics (16%) and fitting (8%).

Of those organisations employing more than one thousand employees 91,7% utilised psychometric tests as criteria taken into account when considering employing apprentices.

Holburn (1989) determined that the most frequently mentioned psychometric tests used to select apprentices are the Mechanical Comprehension test (45,7%), Blox (35,7%), Mental Alertness (20,9%) and Figure Classification test (15,5%).

After evaluating the psychometric tests, Holburn (1989) recommended that they be evaluated for their suitability in multi-cultural applications, and suggested that trainability tests be considered for incorporation in the selection programmes. It was also recommended that consideration be given to evaluating the psychometric tests most frequently used in terms of their bias content, which gave rise to a later study by the same author.

Hughes (1989) conducted a study to ascertain whether the psychometric test used in the election of apprentices discriminated unfairly against any race group. Data was attained for 177 white and 157 black males at a mining training centre in the Transvaal. Of these, 41.8% of the white apprentices were electrical apprentices and 53.5% of blacks.

The predictors which were used were not the same as in the current study, but the domains of competence are similar. The trade test result and technical college results were used as criteria. The psychometric tests used were: (1) the Otis test, a mental alertness test measuring general intelligence. It comprises of items in the form of vocabulary, analogies, alphabetic coding, series continuation, verbally phrased arithmetic problems, and relationships (similar to the Mental Alertness). (2) Number Series Test (NST), a test of general reasoning by means of determining the relationship between the numbers in a series, and then completing the series. (3) Space Perception Test (SPT), a test which measures the ability to perceive and mentally rotate a two-dimensional representation (similar in concept to the Blox) (Hughes,1989).

It was hypothesized that "due to the enforced separate political, socio-economic, and educational lifestyles of the two groups in South Africa, that the psychometric tests used by the company would unfairly discriminate against the black group." (Hughes, 1989:168). It was anticipated therefore, that the predictor scores would be lower for the black group than the white group but the criterion scores would be equal or greater, thus indicating that the black group were as competent as the white group in the trade test and academic situation. This was proved to be correct, suggesting that the three

psychometric tests tended to under-predict black apprentices criterion scores, indicating unfairness to the black group.

The decline in the percentage of apprentices qualifying as artisans led Horn (1990) to evaluate a psychometric test battery's appropriateness which was being used by a company in heavy industry. The battery consisted of: (1) Intermediate Battery Mental Alertness, (2) Intermediate Battery Arithmetic Problems, and (3) the Blox test.

Fifty three apprentices (20 white and 33 coloured) were used in the sample. The primary criterion in the study was the qualification of apprentices as artisans at the apprenticeship examination. The secondary criterion was academic success at technical college. Horn (1990) found that the correlation between the psychometric tests and the academic results indicated that only the Mental Alertness test and the technical drawing scores at college produced a statistically significant correlation, and this was for the white group only. The correlations between the psychometric tests and the primary criterion were statistically insignificant for all groups.

O'Connor Harrison (1972) conducted an analysis on a sample of 192 white apprentices employed at the South African Railways. A selection battery comprising of six psychometric tests, which included the Mental Alertness and Blox test was correlated against first year technical college results and practical assessments. Results were that the Mental Alertness had a significant correlation with both the theoretical and practical criteria, and that the Blox test proved to have a significant relationship with all the criteria except two.

Wheeler (1993) assessed a sample of 93 black male apprentices (comprising of 24 electical apprentices) and 166 white males apprentices (57 electrical) employed at three mines in the Transvaal. The psychometric test battery comprised of five tests, including the Mental Alertness and Blox tests. The criteria against which the test scores were correlated were technical college results and job performance via a performance appraisal system.

The findings of the study were that there were significant differences in means between the black and white groups on all the predictors. A t-test was performed on the predictors with the result that the Mental Alertness test showed $t = 4.39$ ($p<0.001$) and the Blox test showed $t = 7.68$ ($p<0.001$). The mean scores indicate that the "black apprentices tended to achieve lower scores on the predictors, in most cases between a half to one standard deviation less than the whites on the tests" (Wheeler, 1993:67).

On the more objective of the criteria, the technical college results, the black sample scored significantly higher than the white sample. Wheeler (1993) also determined that the measures with the highest predictive validity for the total sample were those which measured the capacity for spatial cognition and abstract thinking, namely the Blox test.

Theron and Barlow-Jones (1979) at the Aptitude Test Centre (Welkom) conducted an investigation into the relationship between aptitude test results and trade test results of apprentice artisans in the mechanical and electrical engineering fields. The objective of the study was to establish the extent to which aptitude test results predicted the success of apprentices in their relevant trade tests.

To a large extent, the aims of this study overlap with those aforementioned, and the methodology used and conclusions drawn by Theron and Barlow-Jones are especially pertinent.

The sample group consisted of 165 apprentices for which both aptitude and corresponding trade test scores were available. In addition, the entire group of 165 had the opportunity to do the trade test regardless what score they had achieved on the aptitude tests. This effectively allowed the researchers to see whether any candidates filtered out by the aptitude test battery could nevertheless have gone on to pass the trade test (i.e. detecting "false negatives"), an element which is lacking in the present investigation. An important point to bear in mind is that this sample was a homogenous group, comprising of white male apprentice.

The psychometric tests used for selection purposes were part of the Intermediate Battery, designed by the National Institute of Personnel Research (N.I.P.R.), being:

* Mental Alertness
* Computation
* Mechanical Comprehension

The criterion against which aptitude test scores were measured was whether an applicant passed or failed the trade test on the first attempt.

One of the statistical methods used to analyse the data was an efficiency ratio (how efficiently the aptitude test battery predicted). This was expressed in the form of a percentage and is derived from the following formula:

Efficiency ratio = No. of candidates above stanine 6 who passed the trade test + No. of candidates below stanine 6 who failed the trade test

Total number of subjects

The figure achieved was then converted to a percentage figure.

The correlation coefficients of aptitude test results against trade test results were the following. The Mental Alertness test had a correlation coefficient of 0,2179 (significant at 5% level) and the Mechanical Comprehension test a correlation coefficient of 0,2279 (significant at 5% level) (Theron and Barlow-Jones, 1979).

Correlations between the aptitude test and trade test scores were all low but were significant. It was concluded that the type of tests used by the Aptitude Test Centre appeared to be the correct tests with which to screen apprentice applicants. However, analysis of the sample of 165 also yielded the following:

Four percent of the sample obtained a stanine of 6 or above on each of the aptitude tests but still failed the trade test (this denotes a false positive prediction). Twenty five percent of the sample obtained a stanine of 6 or more on each of the aptitude and also passed the trade test (denoting a true positive prediction). Twenty four percent did not obtain a stanine of 6 and also failed the trade test (a true positive). Forty seven percent did not

obtain a stanine of 6 but passed the trade test (a false negative) (Theron and Barlow-Jones, 1979).

From the above it appears that the aptitude test results were incorrect in leading to the acceptance of six applicants into training who were not able to pass the trade test on their first attempt, as well as predicting that seventy seven applicants would fail the trade test when in fact they did not. The efficiency ratio computed from the above data resulted in the following:

Efficiency ratio=[(42 + 40) / 165] X 100 = **49,7%**

It was decided that the efficiency ratio was unsatisfactory and the reason proposed for such a dismal efficiency was that the stanine cut-off of 6 was unrealistic. The alternative arrived at was that instead of insisting that the applicants achieve a stanine of 6 on every test, the sum total of the stanines should be 18, based on the principle that a low score on one test could be compensated by a high score on another. This principle applied resulted in the efficiency ratio being improved to 66,6%.

Instead of merely adjusting the stanine criteria for acceptance, the predictive validity of the tests themselves could have been scrutinised, and an alternative to the tests might have been investigated. There are numerous adjustments that could be applied to aptitude test scores, such as increasing cut-off stanine scores, considering the stanines as a whole by averaging them out, etc. Once data manipulation becomes the method used in order to improve test effectiveness, the danger of losing sight of the real goal, to ensure the most appropriate test is used for the purpose, becomes a concern.

The correlation coefficients achieved in the study were not sufficiently high enough for the test user to feel confident that the relationship between predictor and criterion was a strong one, albeit that they were statistically significant. Furthermore, the fact that the aptitude tests, when constructed, were standardised on the same population group as the 165 that wrote them makes the strength of the correlations (the highest for an individual test was 0,2279) all the more concerning. With such low correlations the researchers should perhaps have begun to investigate alternative methods for selection.

Summary

Previous research in the field of validation of test batteries for apprentices has yielded varying results. Generally, the results yielded by the studies do not show a strong positive relationship between the predictor scores and the criteria.

Hughes (1989) found that blacks scored significantly lower on the predictors than whites, but on the criterion scores the results were equal to the white sample or greater. Wheeler (1993) concurred with this assessment. Horn (1990) found the correlation between the psychometric test and the criteria to be insignificant for all groups. O'Connor Harrison (1972) found the Mental Alertness and the Blox test to correlate significantly, and Theron and Barlow-Jones (1979) found the Mental Alertness and Mechanical Comprehension test to correlate significantly but not strongly with the criteria.

Holburn (1989) captures the overall sentiment regarding the psychometric test battery's appropriateness for apprentice selection by recommending that the tests be evaluated for multi-cultural applications and that the alternative of trainability testing be explored.

The conclusions drawn by the above mentioned researchers will provide a useful contextual backdrop to the interpretation of the data in chapter seven.

CHAPTER SIX

6.0 Test bias in the predictors

The issue of bias and fairness of the selection battery needs to be raised, because it is these very concepts that will be questioned by parties declaring a case of unfair discrimination. The question of bias in the tests also needs to be examined in order to keep in mind when assessing the results of the statistical analysis in the next chapter. The definitions of bias and fairness are complex and varied.

Verster (1985) notes that bias can be considered present in psychological testing when scores are differentially, but systematically influenced by sources irrelevant to the construct being assessed. Potential sources of bias include item format, item content, test language, etc. (Hughes, 1989).

Fairness in psychometric testing has to do with the quality of the decision for selecting one individual rather than another (Verster, 1985). Fairness is distinguishable from bias in that it concerns the use of the test scores after they have been obtained. Bias has to do with influences on the test scores during testing (Hughes, 1989).

It is important to acknowledge that a test battery is designed to discriminate. A valid selection measure accurately discriminates between those with high, and those with low probabilities for success on the job (Cascio, 1987). The issue is whether the test discriminates fairly. Guion (1966:26) states that unfair discrimination exists when

"persons with equal probabilities of success on the job have unequal probabilities of being hired for the job".

Hughes (1989) states that, "at the heart of the question of test fairness is the question of validity. Of particular relevance in personnel selection is criterion related validity".

The only study which comprehensively addresses the question of test bias and fairness with regard to the four psychometric tests used by the Department of Electricity is that conducted by Holburn (1992).

Holburn (1992) undertook to investigate the suitability of Intermediate Battery Mental Alertness, the Mechanical Comprehension, the Blox, and the High Level Figure Classification Test for various racial groups in industrial settings. Her findings are vital for the purposes of this investigation.

Holburn's study focused initially on a content validation of the four tests for different race groups, and the conclusions arrived at in terms of their content validity are very important in terms of providing an explanation for the results of the predictive validation as well as being a useful means of providing context for the examination of the tests. Holburn encountered similar problems to those endured in the process of this study in that diffuculty was experienced in finding adequate sample sizes, especially those for the black population group, quite simply because at the present moment the amount of black individuals entering the trades is still not as great as those from the white population group. The scenario in preceding years was even more unbalanced. As a result, Holburn

conceded that due to limited sample sizes the findings of the study should viewed cautiously.

The use of these particular four tests was seen as particularly problematic when a group of examinees comprised of members of different racial groups.

As a result of legislation in the past, certain racial groups were expressly prohibited from entering specific job categories, and resultingly psychometric selection tests designed to be used for selecting individuals into those fields were developed for and standardised on whites only. Many of those tests are currently still in use today.

As more people from previously educationally marginalised population groups become absorbed into the trades, it is important that the cross-cultural comparability of the test scores be determined in order to ascertain their suitability for use in a multicultural context (Holburn, 1992).

Secondly, the use of a common test for all applicants regardless of race or culture assumes that the candidates have all been exposed to the same educational and developmental opportunities. However, the South African arena was characterised by a large degree of variation in both the quality of education as well as exposure for cognitive development, and as such it is important to ensure that whatever psychometric assessment methods are used, cognisance is taken of this fact.

The objective of the investigation was to empirically examine item and predictive bias in four tests which are frequently used and generally accepted as being the standard test battery for apprentice selection in the trades. Item and predictive bias was done through intergroup comparison.

Item bias was understood as the extent to which a test item functions differently for different population groups. In other words, an item is said to be biased if, the members of Group 1 obtain an average score on that test item which differs from Group 2 by more or less than would have been expected based on Group 1's performance on other items of the test (Taylor, 1987).

The sample selected comprised of applicants for apprentice positions in three South African companies and were drawn from all racial groups. The black-white intergroup comparisons were especially interesting and relevant because it is this comparison in test scores and predictive validity which will be focused on later.

6.1 Item bias in the predictors

Mean score differences between race groups on all four tests were observed, with the largest differences in mean score occurring between the white and black groups.

On the Intermediate Battery Mental Alertness Test, a considerable amount of item bias was found when the black and white samples were compared. Seven out of the 30 test items emerged as being biased against the black participants. The test items found to be

biased were item 15, 16, 20, 23, 25, 27 and 28. Most of the biased items were of the alphabetic type (that is, based on the alphabet), and requires the applicant to decipher or decode alphabetic codes or complete alphabetic series. An example of a typical sort of question is given. (This is not an actual test item):

"If **dog** is coded as **eph**, *cat* would be written as ___"

When the biased items were removed from the test and the test rescored, the black-white mean difference narrowed considerably, the test performance of the black candidates improving relative to the white group. Possible reasons for these items - specifically those involving alphabetic interpretation - resulting in large black-white mean score differences will be explored at a later stage.

6.1.1 Mental Alertness test

Holburn (1992) proposes a method for dealing with the item bias in these tests. In the case of the Mental Alertness, where almost one third of the test items appeared to be biased against blacks, some adjustment or correction to the scores seems warranted if bias is to be eradicated from the test.

It is suggested that the test score for black examinees be calculated from the unbiased items only. The test has 30 items of which 7 are proved to be biased, therefore the test would be scored out of the 23 items that are not biased for the black group. The total score out of 23 would be multiplied by 30 and divided by 23 in order to equate the scores

with the other (white) group which is scored out of 30. A second option would be to eradicate the biased items from both black and white groups, and score both groups out of the remaining 23 items.

Alternatively, separate norms could be used for the different groups (as is presently done). However, although the use of separate norms would compensate for different group scores, it will not provide an exact correction for bias. The charge could also be put that using separate norms merely reinforces the notion of separate standards and does nothing to correct the root of the problem and level the playing fields.

6.1.2 High Level Figure Classification test

The High Level Figure Classification Test has been shown through previous research to be one of the most suitable test to use on all examinees regardless of racial breakdown (Holburn, 1992). The amount of bias emerging from this test was negligible, with only one item (item 17) found to be biased against blacks. The item found to be biased appeared to be of a different type compared to the rest, having a three-dimensional visual-perceptual component to it, which the other items do not have.

Removing the biased item and rescoring the test did not alter the mean difference in scores between the black and white groups. Holburn therefore concluded that as a result of the small amount of item bias present in the test and the trivial effect it has on test performance, the test could probably be used in its present form for all groups.

6.1.3 Blox test

In analysing the Blox Test, 3 out of the 45 items were identified as biased against blacks in the black-white comparisons (items 11, 12 and 24), although the reasons for the bias were not clear. It was recommended that the same correction method be used as with the Mental Alertness Test.

6.1.4 Mechanical Comprehension test

According to Holburn (1992) it has been continually revealed through experience that the Mechanical Comprehension Test does not operate well and is inappropriate for use with black groups.

More items than on any other test were found to be biased against blacks and on this basis Holburn recommends that this test should not be used for multicultural selection. The layman looking at the test questions asked in the Mechanical Comprehension Test might be able to identify possible problem with bias from a quick review of the test. Many of the questions rely on a Std 8 (Grade 10) equivalent of general science and also take for granted a fair amount of personal experience.

For example a question dealing with sound and its travel through mediums uses the concept of the 'tin can telephone' - two cans attached with a piece of string as the basis for a question. Any candidate who as a child made and played with such a device would probably find such a question easy to answer, but those who have never seen or been told about this type of thing would be completely ignorant. There are more questions that

deal with light and acoustics which also take for granted an element of general knowledge which could only be attained by exposure to that type of situation.

6.2 Predictive bias results

A predictive bias analysis was also undertaken in the investigation with the aim of investigating the tests in terms of them being similar predictors of criterion measures for different groups. The criterion measures used were monthly progress reports as well as an efficiency report which is an assessment of each apprentice's performance on the training modules.

Once a predictor and a criterion score has been obtained, a regression line can be computed which indicates the linear relationship between the two variables. If a test is a biased predictor for two different groups then the regression line will have a different slope for each group, for example if the predictor scores for the two groups are similar but the criterion scores differ widely. Therefore, when predictive bias is present, the regression lines indicating the linear relationship between the predictor and criterion scores are not the same across the samples.

Because of a lack of available test data, predictive bias analyses could only be conducted for the HL-FCT and the Blox Test.

When test scores for these two tests were used to predict two criteria of first year performance on training modules, predictive bias was detected for the Blox test between

black and white apprentices (Holburn, 1992). Although the black apprentices had obtained on average lower scores on the predictor, they obtained higher job performance scores than their white counterparts. Holburn states that on the Blox test, black apprentices had an average test score which was three standard deviations lower than that of white apprentices, whereas their average job performance score was 4% higher. One method for reducing the predictive bias in this case could be to lower the Blox Test cut-off acceptance score for the black applicants, possibly by one stanine point, in order to allow those applicants who would not make the acceptance criteria, the opportunity to prove themselves on the training course. If the precedent of criterion scores achieved by previous black apprentices are any indication of the potential of future groups, then on the job performance should improve relative to scores on the Blox Test.

For the High Level Figure Classification Test, no significant predictive bias was observed.

Summary

The issue of the presence of bias in the four psychometric tests has been adequately examined by Holburn (1992). Unfortunately there not been other research carried out in the same field which has such direct relevance to the present study.

It was found that in terms of item bias, the Mental Alertness and Mechanical Comprehension tests showed considerable bias against the black sample. Techniques were suggested in order to overcome the effects of this, but it would appear that these tests are affected to such an extent that they should be considered inappropriate for use in

a multicultural situation. The reason for such a large amount of bias could be construed to be a result of the verbal content of the tests and their reliance on alphabetic interpretation in the case of the Mental Alertness test, and previously learned knowledge in the Mechanical Comprehension test. These two tests could be termed measures of "crystallized intelligence".

The High level Figure Classification test and the Blox test showed negligible item bias, which may be put down to the fact that they are non-verbal in content and measure abstract reasoning or "fluid intelligence".

In terms of predictive bias, data was only obtained for two tests. The High Level Figure Classification test showed no predictive bias, while bias was predicted in the Blox test. On this test, the black sample had scored significantly below the white sample, but this was not duplicated on the criterion measure. The Blox test underpredicted for the black sample group, corroborating with the studies carried out by Wheeler (1993) and Hughes (1989).

From the previous two chapters, the predictive validity and appropriateness of the psychometric test battery used by the Department of Electricity has been seriously questioned. The next chapter will analyse and discuss the results obtained from this study.

CHAPTER SEVEN

7.0 Methodology

The data gathering technique used and results obtained will now be discussed. The most significant problem experienced was the lack of both predictor and criterion scores availability for all candidates. Although there was a large body of data, it was only of worth to the study if both assessment results were available for each apprentice. Secondly, the disproportionate amount of black and white apprentices results meant that only a very small black sample was available, and the problems associated with small sample sizes became a concern. The sample consisted of a total of 55 apprentices, 16 black and 39 white.

Both raw and normed scores were obtained for each candidate, as well as a trade test score, represented as a percentage. Raw scores were converted to stanines according to norm tables for each psychometric test.

The aim was to conduct a correlation analysis of the data in order to determine if there was a significant link between predictor scores and trade test results. Criterion-related predictive validity was the central concern of the analysis. The central issue was how effectively the psychometric test battery was predicting the apprentices' performance on the trade test. In order to establish the psychometric battery's predictive power in this context, there must be an empirically demonstrated correspondence between the assessment criteria. A correlation analysis provides a direct check on how well the

predictor is measuring what it purports to measure by providing an indication of the strength of the relationship between the two variables.

The data was entered into a spreadsheet format (Excel software), with the predictor scores corresponding to the trade test score. As a result of the differential results between black and white samples in similar investigations, the unconfirmed assumption was that the predictor and criterion scores would be different for the black and white candidates. The data was therefore coded by fixing a number to members of the each race groups, "1" for blacks and "2" for whites. This was done in order that a t-test analysis could be performed on the sample to ascertain if the black group had scored differently to the white group. The data was then imported into a statistical software package (SPSS) and analysed.

A t-test analysis was firstly conducted on the normed score of each group per psychometric test as well as the trade test. The result was that the samples were significantly different on three of the psychometric tests. It was decided that the samples would be analysed separately as a result. A correlation analysis was then performed on each sample separately.

7.1 Analysis of results

In this section the results of the analysis of the data will be conducted. Normed score were used to conduct the t-test because it is the normed score and not the raw score

which is actually used in the selection. In the following table the mean score for each sample, blacks and whites is shown as well as the t-value.

Table 1.0

T – test analysis for black and white group (mean values)

	Black	**White**	**t value**
HL-FCT	8.00 (n = 15)	6.92 (n = 25)	3.15 **
Mental Alertness	7.42 (n = 14)	6.64 (n = 39)	1.48
Mechanical Com	7.31 (n = 16)	6.28 (n = 39)	2.31 *
Blox	7.87 (n = 15)	6.16 (n = 39)	3.94 **
Trade test	64.88 (n = 16)	71.49 (n = 39)	-1.41

* $p<0.05$

** $p<0.01$

From table 1.0 it is clear that in terms of the stanine mean for each of the four psychometric tests, the black group scored higher than the white group. However, on the trade test , the black group scored below the white group. The inference of this is that the psychometric tests are over predicting the performance of blacks on the trade test, a phenomenon which is in contrast to all other research on this matter.

On the HL-FCT and Blox tests, the difference in mean scores between the two samples was significant at the 1% level, and on the Mechanical Comprehension test the difference in means was significant at the 5% level.

As a result of the t-test, the black and white sample were assessed separately as it was clear that the two groups were behaving differently on the assessment criteria.

A correlation analysis between the predictor score and the criterion was then carried out for each of the two samples.

Table 1.1

Correlation between predictor score stanines and trade test (blacks)

	r	p	n
HL-FCT	0.156	0.577	15
Mental Alertness	0.383	0.176	14
Mechanical Com	-0.01	0.997	16
Blox	0.227	0.414	15

Table 1.1 shows that there was no significant correlation between any of the psychometric test scores and trade test results for the black sample group. The correlation co-efficients are very small apart from the Mental Alertness test, and none are significant. For the Mechanical Comprehension test a negative relationship exists. None of the psychometric tests have any predictive validity for the trade test.

Table 1.2

Correlation between predictor score stanines and trade test (whites)

	r	p	n
HL-FCT	0.232	0.263	25
Mechanical Com	0.12	0.453	39
Mental Alertness	0.38	0.017*	39
Blox	0.11	0.485	39

Figure 1.2 indicates that the HL-FCT, Mechanical Comprehension and Blox tests do not have a significant correlation with the trade test. The Mental Alertness test however, does have a fairly strong positive correlation with the trade test ($r = 0.38$) and is significant at the 1% level. It is therefore the only test with any predictive validity for the trade test.

Finally, a correlation analysis between the raw scores and the trade test for bot groups was conducted. Raw scores provide an "unfiltered", and "uncorrected" indication of the scores achieved by a candidate on the tests before norming. Converting raw scores to stanines essentially provides a compensation factor to black scores by adjusting the scores so as to be on more level ground with white scores. The raw scores were included in the study in order to see if they provided the same result as the normed scores.

Table 1.3

Correlation between raw predictor scores and trade test (blacks)

	r	p	n
HL-FCT	0.23	0.39	15
Mechanical Com	0.15	0.57	16
Mental Alertness	0.19	0.44	14
Blox	0.13	0.31	15

The results achieved when applying the raw scores mirror the results of the stanines. Correlation co-efficients are still very small and no relationships are significant.

Table 1.4

Correlation between raw predictor scores and trade test (whites)

	r	p	n
HL-FCT	0.13	0.53	25
Mechanical Com	0.10	0.51	39
Mental Alertness	0.35	0.02*	39
Blox	0.14	0.37	39

* $p<0.05$

Table 1.4 reflects much the same as table 1.2. The Mental Alertness is the only psychometric test which shows a significant positive relationship with the trade test.

In summary, the psychometric test battery has been proven to be a weak predictor of performance on the trade test for both sample groups. The only test which has predictive validity is the Mental Alertness test, but for the white group only.

It would have been worthwhile to assess the psychometric tests with the test items which were identified as biased by Holburn excluded. However, this was not possible as only the test scores were available, not the written tests themselves.

There are some concerns regarding the results achieved. Because of certain limiting factors, the findings must be regarded as tentative and viewed with caution.

7.2 Limitations of the study

There are a number of factors which will limit the generalizability of the findings of the study.

7.2.1 Size of the sample

The first concern is the size of the sample that was used in the study. Obviously data could only be utilised if both the original psychometric test and trade test scores were available. It was often the case that although one set of scores were available, records of the other was unobtainable for one reason or another. The size of the sample was therefore as large as could be made possible, and still yields useful information, but yet is still not large enough to draw conclusive results that the researcher could be comfortably confident about.

7.2.2 Restriction of range

The fact that only those who had been selected for training on the basis of their psychometric test results could be used in the study leads to a phenomenon known as restriction of range. This can be explained as follows:

If all the psychometric test results of the initial applicant sample could be plotted on a graph, the scores would tend to conform to a bell-shaped curve known as a normal distribution. Those in the upper percentiles i.e. lying on the far right side of the graph, would be in contention for selection, while those falling below the cut-off point would be rejected. The result is that this study utilises only a very select group of applicants (those in the upper quartile), and this has a direct distorting effect on the statistics generated from the sample. Schepers (1992:1) states that,

"In validating tests for the use of personnel selection, special care should be taken to ensure that the obtained validities are neither underestimates not overestimates of the true validities. The following (factor) merits special attention in this regard: the degree of self-selection that has taken place prior to the onset of the study."

Schepers proposed a mathematical formula for the correction of restriction of range. This formula will not be applied to the data due to the complexity of calculations involved. The results generated will still be useful, although statistically 'raw'. A statistician's comment regarding the sample size and restriction of range problem was, "It is not so much a statistical problem as a logistical one!" (Piper, 1996).

7.2.3 Lack of a control group

It would be of great value to this exercise (albeit impossible under the circumstances) if every applicant that was psychometrically tested was also put through the training program and allowed to do the trade test regardless of their psychometric performance. If the psychometric tests are predicting trade test performance accurately, all those who fell below the cut-off limit on the psychometric tests should logically fail the trade test. Conversely, all those who scored above the psychometric cut-off should pass the trade test.

If this was proven so, then the predictive powers of the psychometric tests in gauging trade test performance would gain much more credibility. Essentially, the results generated, because they are based on the data available are showing only half of the potential picture. It is obvious that the psychometric tests are not predicting perfectly because some of those selected actually failed the trade test. These are the 'false positives'. It would be enlightening to investigate how many 'false negatives' there would be out of the sample, that is, would any psychometrically unsuccessful applicants have been able to pass the trade test if given the training opportunities which the successful applicants had?

Above are some of the factors inhibiting the statistical accuracy and generalizability of the results of the study. They do not invalidate the results, but are important to take into account when interpreting them.

7.3 Conclusion and recommendations

The results of the validation study have identified the shortcomings of applying what may be termed "traditional tests" to a changing South African society. Some of the shortcomings of conventional testing are that there has been a belief that culturally determined skills actually measure "intelligence". The case of the Mental Alertness and Mechanical Comprehension tests identify this. Two of the four tests employed measure what could be termed "crystallized intelligence" are do not tap into the individual's ability to learn or assimilate new information. The tests essentially measure static skills, accumulated historically as the individual is socialised. In other words, they measure competencies that have been acquired in the past. There is no focus on what could be developed in the future.

The philosophy underpinning the Mental Alertness and Mechanical Comprehension test, for example, takes its bearing from early psychometric theories in that specific abilities are isolated and measured. In addition, these tests are verbally oriented and assess previously gained knowledge, rendering them patently culturally loaded to the observer.

A second shortcoming is that the abilities measured by the conventional tests are often not relevant to the actual competencies required for the job. Although this study has not focussed on the construct validity of the tests, through the identification of competencies by way of observing the trade test pilot runs, there did not appear to be much overlap between competencies required by the predictors and those required by the criterion, except for the spatial component. The measuring of competencies irrelevant to the job

would, as has been discussed, expose an employer to possible unfair labour practice cases.

It might be recommended therefore, taking cognisance of the past history of testing, the pressures being placed on business and industry to adapt to a multicultural environment which is free of bias or unfair advantage, and taking into account the findings that have emanated from the investigation, that alternative assessment procedures be sought by the Department of Electricity.

It is also recommended that an instrument designed to assess future potential for learning and development, especially in the competencies of inductive and deductive reasoning, (as identified through observation) be sought. There are a number of test developers that have explored this idea.

Taylor (1994) has sought to find a way of going about assessment that addresses the concerns facing South Africa. He attempted to find ways to identify those who "have potential for development, even though they may have gaps or limitations in their skill repertoire due to past disadvantagement." (Taylor, 1994:191)

To achieve this aim, he designed a test battery which would assess potential rather than 'crystallized skills' or specific abilities, which he called the APIL-B (Abilities, Processing-Information and Learning Battery). He attempted to "avoid the numerical and verbal domains as much as possible as competence in these is very much a function of quality and quantity of schooling." (Ibid).

The APIL-B utilises abstract-diagrammatic material in its tests which was designed to be acultural and free of context, and will be material that the candidate has never had to deal with before. Essentially, the battery measures performance through differenet stages. Evaluation starts off from a base level, where the candidate is still trying to come to grips with the new material. The candidate is then exposed to learning opportunites and is retested on the same material later. The improvement in the candidate's performing, or the 'curve of learning' which results is the indication of the candidate's ability to learn and master new material and to learn a new competency.

The above test battery is certainly not suggested for the purpose of selecting apprentice electricians, but merely explains an example of the direction the City Council should be heading towards if they are to achieve an unbiased, useful, and valid assessment instrument for the selection of apprentice electricians. There are a number of new tests that approach the problem of testing in the same way.

It has been shown in this study that current psychometric tests are fraught with problems and weaknesses. From the roots of psychometric theory and the very ideas underpinning psychological testing certain difficulties were bound to arise, and these have become even more problematic considering the present South African context. Research relating to the use of the four tests investigated has been contradictory and varied. However, there has never been an authoritative study that has concluded that the tests are valid, fair for all race groups, useful for apprentice selection and therefore legally "watertight" for the South African situation. The results of this investigation extend on the findings of previous researchers that, at best, the tests have limited applicability for the selection of

apprentice electricians. In line with current thinking and theories around test development and test use, it would be advisable that alternative means for the selection of apprentices be found, especially considering the proliferation of options available.

REFERENCES

Aiken, L.R. (1982). Psychological testing and assessment .(4th Ed). Allyn & Bacon

Anastasi, A. (1982). Psychological testing. (5th Ed). New York: MacMillan

Arvey, R.D. & Faley, R.H. (1988). Fairness in selecting employees . Reading, MA: Addison Wesley

Barrett, G.V. & Bass, B.M. (1975) Cross-cultural issues in indistrial and organisational psychology. In M.D. Dunnette (Ed.). Handbook of industrial and organisational psychology. Chicago: Rand McNally

Biesheuvel, S. (1949). The selection of engineers. Engineer and Foundryman. Johannesburg: Rosta Printers

Campbell, M. (1995). Interpretation of psychometric test scores as a factor leading to poor test scores among Zulu speakers, with special reference to the Higher Level Mental Alertness Test .(MA thesis) University of Natal, Durban

Cascio, W.F. (1987). Applied Psychology in Personnel Management. (3rd ed.) Englewood Cliffs: Prentice-Hall

Cole, N.S. (1981). Bias in testing. American Psychologist, 36 (10), 1067-1077

Cronbach, L.J. (1949). Essentials of psychological testing (3rd Ed). Harper and Row

Dreger.R.M. & Miller, K.S. (1960). Comparative psychological studies of negros and whites in the United States. Psychological Bulletin, 57, 361-402

Dreger,R.M. & Miller, K.S. (1968). Comparative psychological studies of negros and whites in the United States. Psychological Bulletin, 70, 1-58

Du Toit, D.G. (1992). Die klassifikasie van vakleerlinge in die metaalnywerheid in spesifieke ambagte met behulp van psigometriese toetse, Master thesis, Randse Afrikaanse Universiteit

Erasmus, P.F. (1995). Standardisation, validity, and reliability of assessment tools in a fast changing culturally diverse society. ITO Focus cc.

Ghiselli, E.E. (1964). Theories of psychological measurement. McGraw Hill Series in psychology

Guilford, J.P. & Fruchter, B. (1978). Fundamental statistics in psychology and education. McGraw-Hill Kogakusha Ltd

Guion, R.M. & Gibson,W.M. (1988). Personnel selection and placement. Annual Review of Psychology, 39. 349-374

Halstead, M.E. & Du Toit, A. (1983). Blox Test: test administrators manual (revised edition). K7.66 Johannesburg: National Institute for Personnel research, Council for Scientific and Industrial Research

Holburn, P.T. (1992). Test bias in the Intermediate Mental Alertness, Mechanical Comprehension, Blox and Higher Level Figure Classification Tests. An NTB, HSRC report, Pretoria

Holburn, P.T. (1989). Apprentice selection: An HSRC/NTB survey of policies and methods used in the RSA with an emphasis on psychometric testing. Contract report C/PERS 406. Pretoria: HSRC

Horn,J. (1990) Die voorspellingsgeldigheid van 'n psigometriese toetsbattery vir keuring van 'n heterogene groep vakleerlinge. Masters thesis, University of Stellenbosch

Hughes, C.E. (1989) The comparison and evaluation of various models of test fairness on samples of white and black employees, MA thesis. University of Port Elizabeth

Intermediate Battery Test Administrators Manual. (1973). Johannesburg: The National Institute of Personnel Research, Human Sciences Research Council

Kail, R. & Pellegrino, J.W. (1985). Human intelligence: perspectives and prospects. New York: W. H. Freeman & Co.

Magnusson, D. (1967). Test theory reading. Massachusetts: Addison-Wesley publishing

Maloney. M.P. & Ward, M.P. (1976). Psychological assessment: a conceptual approach. New York: Oxford University Press

Mechanical Comprehension Test Administrators Manual (1968). Johannesburg. The National Institute of Personnel Research, Human Sciences Research Council

Michael, W.B., Guilford, J.P., Fruchter, B., & Zimmerman, W.S. (1957)The description of spatial-ability abilities. Educational and Psychological Measurement, 17. 185-199

Minister of Labour (1998). Employment Equity Bill. As introduced to the National Assembly. B 60-98

Moss, P.A. (1992). Shifting conceptions of validity in educational measurement: implications for performance assessment. Review of educational research volume 62, no 3. University of Michigan

National Union of Metal Workers of South Africa (NUMSA) (1992). Code of practice to end unfair discrimination in employment practices. Unpublished manuscript.

Nunally, J.C. (1970) Introduction to psychological measurement. McGraw-Hill Book Company

O'Connor Harrison, B.A. (1972). Die voorspelling van sukses van vakleerlingepassers gedurende die opleidingstydperk. MA Thesis, University of South Africa

Pieterse,C. & Bowden, G. (1993). Why do our apprentices fail? In "Electrical Contractor": Nov/Dec

Piper, S.E. (1994). Postgraduate research methods course notes. University of Natal, Durban

Poortinga, Y.H. (1983). Psycholgical approaches to intergroup comparison. The problem of equivalence in South Africa. In S.H. Irvine & J.W. Berry (Eds.). Human Assessment and Cultural Factors. New York: Plenum Press

Retief, A. (1988). Method and theory in cross-cultural psychological assessment. HSRC Research Report Series No 6. Pretoria: Human Sciences Research Council

Schepers, J.M. (1994). The development of a statistical procedure to correct the effects of restriction of range on validity coefficients. Department of Human Resource Management, Rand Afrikaans University

Society for Industrial Psychology (1992). Guidelines for the validation and use of personnel selection procedures. Pretoria: Society of Industrial Psychology

Spies-Wood, E. (1988). Bias, comparability, fairness, and utility in cross-cultural testing:implications in the South African context (Office Report 1988-09). Pretoria: Human Sciences Research Council

Sternberg, R.J (Ed.).(1985) Human abilities – an information processing approach. New York: W.H. Freeman

Steyn D.W. And Latti V.I. (1974) Die ontwikkeling van 'n strategie vir die keuring van eerstejaarstudente in die fakulteit ingenieurswese van die Universiteit van Pretoria. Johannesburg: Nasionale Instituut vir Personeel Navorsing

Taylor, J.M. (1983, February). The prediction of success of black engineering technicians. Paper presented at Psychological Association of South Africa Annual Congress, Pietermaritzburg.

Taylor, J.M. (1987a). Test bias: the roles and responsibilities of test user and test publisher. Special report: Pers 424, Pretoria: Human Sciences Research Council

Veldsman, T.H. (1990, February). A psychometric strategy for the South African Mining Industry. Paper presented at the Fairness in Personnel Selection Seminar, Johannesburg

Verster, J.M. (1987). Bias and fairness in psychological assessment: clarification of terms. Unpublished manuscript, HSRC workshop on bias and fairness in psychological testing. Johannesburg: NIPR/HSRC

Walsh, W.B. & Betz, N.E. (1985). Tests and assessment. Englewood Cliffs: Prentice-Hall

Wheeler, H.L. (1993). The fairness of an engineering selection battery in the mining industry. MA thesis, University of South Africa