AN INTEGRATIVE REVIEW OF THE NIPR-DEVELOPED INTERMEDIATE MENTAL-ALERTNESS TEST

by

JAMES READER

Submitted in partial fulfilment of the requirements
for the degree of Master of Arts in
the Department of Psychology, University
of Natal, 1991.

PREFACE

This mini-dissertation represents original work by the author. Where use has been made of the work of others, this has duly been acknowledged in the text.

ACKNOWLEDGEMENTS

I would like to express gratitude to the following people whose assistance has been invaluable in the completion of this mini-dissertation:

Ms. Bronwyn Brown, and particularly Ms. Louise Fenske for their assistance in the typing of the text;

Ms. Frances Browne, for the complex task of typing the tables;

Dr. Terry Taylor, of HSRC Johannesburg, for providing useful information in the area of item bias and fairness;

Ms. Penny Holburn, of HSRC Johannesburg, for providing useful information and data concerning item bias research on the IMAT;

Mr. Reg Lombard, the TCRSA Johannesburg, for important historical information concerning the IMAT;

Prof Johan Schepers, of the RAU Psychology Department, for important historical information concerning the IMAT;

Ms. Nanette Tredoux, of HSRC Johannesburg, for providing useful information in the area of the computerisation of the IMAT;

Ms. Lorraine Pratt, of the HSRC library in Johannesburg, for her friendly and efficient help with obtaining the references sources;

Dr Hillary Bennett, of the Department of Industrial and Labour Studies at Durban University, for her administrative assistance;

and very importantly, to my supervisor Mr. Jonathan Taylor, Regional Director of HSRC in Durban, for his patience, time and assistance.

ABSTRACT

In this study a rationale is provided for the need for integrative psychometric studies in addition to exploratory studies. The psychometric concepts of validity, reliability, test comparability and bias are reviewed.Furthermore the concepts of the nomological network and meta-analysis are reviewed with regard to their relevance for integrating findings of psychometric research.

The historical origins of the Intermediate Mental Alertness Test (IMAT) are traced with a view to distinguishing it from a number of related tests. Psychometric information is summarised from a wide cross-section of both published and unpublished studies on the IMAT, and the findings of these studies are reviewed in terms of the reliability and validity of the IMAT across different contexts and groups. It is found that there are an insufficient number of validity studies using similar criteria for the meta-analysis technique to be applied. An attempt is therefore made to interpret trends from the research studies reviewed, and recommendations are provided in terms of reporting standards for future validity studies in order to facilitate integrative research techniques such as meta-analysis.

AN INTEGRATIVE REVIEW OF THE NIPR-DEVELOPED INTERMEDIATE MENTAL ALERTNESS TEST

CHAPTER 1

OVERVIEW

1.1 PURPOSE OF THE STUDY

Currently any reference to "The Mental Alertness Test" is likely to cause considerable confusion. The reason for this is that there are a number of different Mental Alertness tests in existence. Although they are all associated with the National Institute for Personnel Research (NIPR)*, these tests differ in various significant ways. Some are direct modifications of an American developed test - the Otis Quick Scoring Mental Ability Test - while others have been developed de novo by the NIPR. Some exist as separate entities, while others are subtests of test batteries. Furthermore, the Mental Alertness tests occur at several difficulty levels, have both English and Afrikaans translations, and were developed with parallel versions. It will therefore be one of the broader aims of this study to untangle and clarify the similarities and differences of the various Mental Alertness Tests.

A report on apprentice selection in the RSA by Holburn (1989) reveals that the Intermediate Mental Alertness Test (IMAT) is one of the four most frequently used tests for apprentice selection, and is therefore an important test in industry.Furthermore,the IMAT is one of the tests used for vocational guidance counselling at the Human Sciences Research Council (Visser,1977),and has been used in a number of other contexts as a predictor,or indicator of cognitive ability (see references provided in Tables 1.1 to 1.11 in Chapter 5). Despite this fact, there has been no comprehensive work to review its reliability and validity performance across different settings and subjects. The current study will attempt to do this.

Bearing in mind the problems of culture loadedness and test comparability (see Chapter 4) it may be useful for test users to know something about how the IMAT was developed and standardised. Officially, the history of the test is lost in the "mists of time". As far as possible then, an attempt will be made to trace the history of the test.

Concerning the standardisation of the IMAT, it is particularly notable that despite revision, the current manual of the IMAT (Wilcocks, 1973), contains only norms obtained from White, mainly male subjects, tested in the mid 1960s. Although it is known that other norms are available through the HSRC, test users may not be aware of the nature and extent of these. This study will therefore include details of norms available through the HSRC.

* (the NIPR was incorporated into the Group: Human Resources of the Human Sciences Research Council (HSRC) as of October 1, 1990)

In general, it is the aim of this study to provide an integrated, concise document on the IMAT which includes:

- information on its history and development
- findings from recent item bias research on the IMAT
- a review of available research on the IMAT in terms of reliability and validity
- recommendations for future research on the IMAT in terms of current research in the area of meta-analysis and validity generalisation

1.2 THE SIGNIFICANCE OF AN INTEGRATIVE STUDY

It is the belief of the author that the integration and analysis of research - drawing the findings of different studies together - is just as important as the creation of new data. This belief is supported by Tukey (1969) who argues that research of an exploratory nature and of a confirmatory nature play equally important roles in the building and expansion of knowledge. He points out that the basis for building a body of valuable knowledge is repetition, which leads to breadth and depth of data building up from parallel sets of research circumstances. Implicit within his argument is that information gleaned at the micro level is brought together to provide a clearer picture at the macro level. "Clarity in the large scale comes from clarity in the medium scale; clarity in the medium scale comes from clarity in the small" (1969 p.87). Attempts to gain clarity at the macro level may be brought about by review-type qualitative research, or by attempts to integrate quantitative data across studies. Recent examples of the former type of research in the field of psychology include Van Staden and Visser (1990), Retief (1986), Biesheuvel (1987), Gardner (1980), Kriegler (1988) and Van Staden (1987).

Concerning the latter type of research, it is a considerable challenge to make sense of large amounts of empirical data. Over the past decade Hunter, Schmidt and their collegues have done much towards cumulating findings across studies. The course of their research has led these authors to develop a statistical technique known as meta-analysis which was devised to enable the cumulation of results across studies, for example Hunter, Schmidt and Jackson (1982); Schmidt and Hunter (1977); Schmidt, Hunter, Pearlman and Shane (1979). On the basis of their meta-analytic research these authors claim that the considerable variability that is often observed amongst essentially similar studies has created the impression of ..."complexity and chaos where in fact there is order and stability" (Schmidt and Hunter, 1980 p.42). Much of the work that Schmidt, Hunter and their colleagues have done has focussed on using professionally developed cognitive ability tests as selection instruments. This work is therefore of particular relevance to the use of the IMAT and will be reviewed later in this study.

CHAPTER 2

HISTORICAL PERSPECTIVE ON THE INTERMEDIATE MENTAL ALERTNESS TEST (IMAT)

The IMAT and Mental Alertness tests were developed within a historical context and a general knowledge of this history is useful in understanding how they came to be developed. This exercise will also help to untangle the distinctions between the different Mental Alertness tests. It may be noted that in the following section reference is made to various tests and batteries that are related to the IMAT, either directly or indirectly. An indication as to how these are related to the IMAT is given in the text and a brief description of the tests and batteries related to the IMAT is provided in Section 2.3 enabling comparisons to be made between these tests.

2.1 THE DEVELOPMENT OF GROUP TESTING AND THE HISTORY OF THE NIPR-DEVELOPED MENTAL ALERTNESS SERIES

While group testing had in a sense been practised in academic settings for generations, prior to the First World War psychological tests had been orientated towards individuals (see Dubois, 1970; Pintner, 1924; Boring, 1950 and Goodenough, 1949, for historical reviews of psychological testing). For example, they often required oral responses from examinees, or manipulation of apparatus, and they usually required a highly trained examiner. The merits of group testing were first set out by Whipple (1910), who pointed out that ".... a class of 50 or 100 children may take a test in less than a fiftieth or a hundredth of the time needed to administer the same test individually" (p. 11).

It was around this time that the idea of presenting questions with the choice of two or more responses emerged. Research psychologists Thorndike and Thurstone were both experimenting with the idea of scoring items with a scoring key (Dubois, 1970). A breakthrough came when psychology research student Arthur Otis developed the first multiple choice, paper-and-pencil, key scored group-intelligence test. It was this test that was to make a significant contribution to the development of group testing, and was the forerunner of the Mental Alertness tests.

At the time of the First World War, the Otis test was as yet unpublished. When the United States entered the war in 1917 the American Psychological Association appointed a committee to consider how psychology could assist the war effort. A decision was made to develop a group test that would be used to place army recruits, according to a set of criteria such as: measuring a wide range of abilities, economy of time and material, ease of scoring, independence of school training, possibility of alternate forms so as to prevent coaching and minimum of writing in making responses (Dubois, 1970). Over a period of a week the committee considered available materials and identified ten test sources for development as subtests for the group test. One of these was the unpublished test of Arthur Otis, which he turned over to them. Additional items were constructed for each of the ten tests, and a prototype group test was compiled called the "Examination a".

This new test was then administered to various experimental subjects ranging from mental retardates to trainee officers to marines. Analysis of performances on the "Examination a" indicated high correlations with a revised version of the Binet, namely the Stanford Binet(Dubois, 1970). Inspired by this, instructions and items were streamlined, poorly performing subtests and items removed, and alternate forms of the tests were printed, culminating in the well known Army Alpha and Army Beta tests. The former was designed for general routine testing, while the latter was a non-verbal test used on illiterates and foreign born recruits who were unable to be tested in English (Anastasi, 1982).

Following the publication of the Army Alpha and Beta tests extensive testing was carried out, involving some 1 726 966 people (Dubois, 1970). The conspicuous success of the program did much to change the perception of psychology as an academic discipline to that of a profession. Towards the end of World War I the publication of the first volume of the Journal of Applied Psychology reflected the growing interest in the application of psychological tools and techniques to the industrial and business sector (Muchinsky, 1987). Following World War I applied testing underwent a tremendous spurt of growth, being applied (often indiscriminately) across a wide variety of people and contexts. Developments in this regard are beyond the scope of this work and may be followed in texts such as Dubois (1970), Goodenough (1949) and Ferguson (1962).

With the advent of World War II, test programs were initiated on an even greater scale than in World War I. A modernised version of the Army Alpha test was developed, the Army General Classification test, and by the end of the war this had been administrated to over nine million people (Huysamen, 1988). In South Africa a need developed in the airforce for a personnel section to handle selections and placements of air crew. This need was met by creating the of the Aptitude Test Section (ATS) of the South African Airforce (Hudson, 1962), which in time not only handled placements, but also aspects of training, stress and morale management, and accident prevention (Biesheuvel, unpublished manuscript).

After the war, in an attempt to promote industrial growth the Board of Trade and Industries recommended that research was urgently needed in industry in order to increase its efficiency and reduce its real costs (Report No. 282, 1945). As a result, the Council for Scientific and Industrial Research (CSIR) was established in 1945. Aware of the work of the ATS during the war years, the founder of the CSIR suggested that this be incorporated into the CSIR to be specifically concerned with the human side of industry. This was agreed upon and the ATS became reconstituted as the Bureau of Research in Industrial Psychology. Over time, the Bureau expanded under the direction of Simon Biesheuvel, and became known as the National Institute of Personnel Research in 1948. More detailed accounts of the development of the NIPR may be found in Hudson (1962), Biesheuvel (Unpublished manuscript), and Huysamen (1988).

The founding of the NIPR provided the opportunity for organisations to utilise the human resource technology that had been developed during the war years. Since the ATS had been so closely associated with testing and selection during the war, the first projects of the NIPR tended to emphasize personnel selection (Hudson, 1962). Initial clients were State Departments such as the Post Office and South African Railways, and various industries (Biesheuvel, 1983). Towards meeting their requests for selection and placement instruments, a psychometric research team headed by A.O.H. Roberts began the task of adapting and standardising wartime tests for industrial application. One of the tests which was chosen for development into an industrial context was the Otis Quick-Scoring Mental Ability Test (OQSMAT - described in Section 2.3 below). This test was a direct descendent of the Otis test which had given rise to the Army Alpha test of World War I, and had been granted to the ATS during World War II for the selection of personnel. Since the OQSMAT had been developed overseas it was necessary that certain modifications were made to adapt the test to local industrial conditions. Consequently Roberts' research team re-worded certain items in the test, and devised new ones, and developed two local versions of the OQSMAT, each with alternate forms (*R. B. Lombard, personal communication, July 1988). Having been standardised for industrial use, these two versions of the OQSMAT became generally known as the Mental Alertness (Advanced) and the Mental Alertness (Intermediate). The alternate forms of the Mental Alertness (Advanced) were catalogued by the NIPR as the A/1/1 and the A/1/2, while the alternate forms of the Mental Alertness (Intermediate) were catalogued as the A/2/1 and the A/2/2. (Further information on these tests is provided in Section 2.3 below).

Over the decade following the founding of the NIPR a need became apparent for a generalisable battery of tests that could be applied across different industrial settings. By the late 1950s a trend was emerging toward extending the use of selection procedures to high level administrative and executive personnel (Hudson, 1962).

At this time, specific requests were arriving from the British governed regions in more northern parts of Africa for instruments to select Black civil servants, and personnel applying for technical positions (Biesheuvel, 1983). In addition, the NIPR was approached by the South African Railways for an instrument to select personnel at the Standard 8 to 10 level of education. Consequently it was decided to develop a battery of tests at different difficulty levels to meet these needs. The items of these tests were to be developed "de novo", but it was decided that they would be of a similar rationale to the Otis tests (**J.M.Schepers,personal communication,July 1988). The work was undertaken by the NIPR psychometrics division under A.O.H. Roberts .

* R B Lombard is a member of the Test Commission of South Africa, and was head of Test Distribution at the NIPR.

** Prof J M Schepers is currently Head of the Department of Psychology at Rand Afrikaans University, and was a member of the Test Division of the NIPR during the time that many of the developments concerning the IMAT took place.

In response to the need for selection instruments for high level administrative and executive positions, items requiring a matric level of education or higher were developed. These items were compiled into a battery of tests that became known as the High Level Battery. In response to the requests from administrators in the Rhodesias (Northern and Southern) and Kenya, as well as local needs for lower level personnel selection instruments, items requiring an education level in the range of Standard 6 to 8 of education were developed (J M Schepers, personal communication, July 1988). These items were compiled into a battery of tests that became known as the Normal Battery. In response to the need of the South African Railways for an instrument to select personnel at the Standard 8 to 10 level of education, the NIPR developed and utilised items that were to become the Intermediate Battery (J M Schepers, personal communication, July 1988). Apart from one or two exceptions (such as the inclusion of a Spot-the-Error subtest in the Intermediate Battery), the subtests of the three levels of battery correspond in all ways except difficulty level. The High Level, Normal and Intermediate Batteries were classified as the A/75, A/76 and A/77 respectively. Each of these was developed in parallel English and Afrikaans versions.

2.2 SCOPE FOR CONFUSION CONCERNING THE INTERMEDIATE BATTERY

In all, three parallel forms of the Intermediate Battery were developed, and these were labelled the A, B and C forms respectively. Each of these forms had Afrikaans and English versions. Once developed by the NIPR, the A and B forms and their different language versions of the Intermediate Battery were made available for general use, while the C form and versions were handed over to the South African Railways for their sole use (R B Lombard, personal communication, July 1988).

Bearing in mind that each of the three forms of the Intermediate Battery had a Mental Alertness subtest, it can be seen that there were three distinct Intermediate Mental Alertness tests, each available in English or Afrikaans. It is important to distinguish these from the Mental Alertness (Intermediate) which was the localised form of the Otis test used by the ATS during the war (see above).

In 1973 there were two significant changes that affected the Intermediate Battery. Firstly, up until that time the catalogue prefix on any given test (for example A/2 or A/77) simply corresponded with the filing system devised by the NIPR. From 1973 the NIPR adopted the American system of classifying tests into A, B and C categories, according to the qualifications of those entitled to use them (see Prinsloo, 1982). It was therefore decided that the catalogue prefix would be changed to indicate the category of test usage. For example, since the Intermediate Battery requires a B level of test user, it was catalogued as the B/77 as opposed to the previously used A/77 (R.B.Lombard, personal communication, July 1988).

In 1973 the second significant change was that the NIPR metricated their tests. When the different forms of the Intermediate Battery in general use were reviewed, it was found that the A form would require an extensive item revision, whereas the B form would require little modification. Consequently, it was decided to abandon the A form of the test (R.B.Lombard,personal communication,July 1988). Therefore, at the time of writing, the NIPR only has one form of the Intermediate Battery available for general use. Categorised as the B/77, this is available in both English and Afrikaans versions .

2.3 DESCRIPTION OF TESTS AND BATTERIES RELATED TO THE IMAT

The last section referred to a number of tests and batteries that are related to the IMAT.This section will provide some more information on these, which will help to differentiate them from each other, and the IMAT.

2.3.1 The Otis Quick-Scoring Mental Ability Test (OQSMAT)

It has already been mentioned that the OQSMAT was used as a model in the development of the NIPR series of Mental Alertness Tests. In fact, three forms of this test were developed, differentiated in terms of difficulty levels as follows:

- Alpha Test — Grades 1-4
- Beta Test — Grades 4-9
- Gamma Test — High Schools and Colleges

Developed in America by A.S. Otis (1937), the three forms of this test were descendants of the earlier tests from which the Army Alpha and Army Beta were derived during World War I. The somewhat ambitiously stated purpose of all three tests is "...to measure mental ability-thinking power or maturity of mind." (Otis, 1937).

The Alpha form is composed of 45 sets of four diagrams of objects or designs. The pupil marks which of the four is the odd one out. The same 45 items may be transformed into a verbal test where the pupil follows directions read by the teacher.

The IMAT is more similar to the Beta and Gamma tests than the Alpha. These particular tests assess the ability of subjects to accurately demonstrate their grasp of opposites, vocabulary, proverbs, similarities (diagrammatic as well as verbal concepts), analogies, number series and reasoning. Both tests have a time limit of 30 minutes, and are of a pencil-and-paper, multiple choice format. A review of these tests is provided by Yates (1959).

2.3.2 The Mental Alertness (Advanced) and (Intermediate)

As mentioned in Section 2.1, the Mental Alertness (Advanced) and (Intermediate) were the direct local derivations of the OQSMAT, which were used in South African industry at the end of Word War 2. These tests were originally classified as the A/1 and A/2 respectively. With the change in the NIPR classification system in 1973, they are now classified as the B/1 and B/2 reflecting their status as B category tests in terms of the Test Commission of RSA classification system. Each having parallel forms, these tests may be differentiated as follows:

- Mental Alertness (Advanced): This test is suitable for testees with 12 years or more of school training.Two parallel forms of the test are available, known as the B/1/1 and B/1/2.

- Mental Alertness (Intermediate): This test is suitable for testees with 10 to 12 years of school training.Two parallel forms of the test are available, known as the B/2/1 and B/2/2.

Being local modifications of the OQSMAT, the B/1 and B/2 have items of a similar nature to this test. Each comprising 60 items, they are of a written answer and multiple choice format and have a time limit of 35 minutes (Van Zyl and Myburgh, 1991).Unlike all of the other Mental Alertness tests, responses to the items of the Mental Alertness (Advanced) and the Mental Alertness (Intermediate) are written in the test booklet, and therefore the test booklets are not re-usable.

2.3.3 The Normal Battery

As was mentioned in Section 2.1, the Normal Battery was developed "de-novo" by the NIPR in response to the need for an instrument to select personnel at a lower level than the target group of the Intermediate Battery. Included in this battery is the Normal Level Mental Alertness Test. Comprising five pencil-and-paper subtests, the Normal Battery was developed in both English and Afrikaans. Four of the subtests follow a multiple choice format, while the fifth (computation) consists of arithmetical problems in which answers are worked out and filled in on the answer sheet. There are 158 items in all, and the full battery takes approximately two and a half hours to administer and complete (Huysamen, 1988). It is intended for those with 7-10 years of formal schooling (Standards 5 to 8).

The English and Afrikaans test manuals provide norm tables covering urban and rural Black scholars and clerks, as well as White scholars, nurses and clerks. Subtests for this battery are as follows:

(a) Mental Alertness (50 minutes)

This is a test of general intelligence comprising 54 items in the form of analogies, alphabetic coding, series continuation, verbally-phrased arithmetical problems, relationships and dissimilarities.

(b) Comprehension (25 minutes)

A test of language ability on the basis of four short paragraphs. Testees are required to answer five questions on the basis of their understanding of the paragraphs.

(c) Vocabulary (15 minutes)

In this test testees are presented with 36 words for which they must choose the correct respective meanings.

(d) Spelling (10 minutes)

In this test 18 words are each spelt in five different ways. From the alternatives, testees choose the correct spelling for each word.

(e) Computation (20 minutes)

Testing subjects arithmetical ability, computation consists of 15 long-division and 15 multiplication problems which must be worked out.

2.3.4 The High Level Battery

As mentioned in Section 2.1, the High Level Battery was developed "de-novo" by the NIPR in response to the need for an instrument to select personnel at a higher level than the target group of the Intermediate Battery. Included in this battery is the High Level Mental Alertness Test. As with the Normal Battery, the High Level Battery occurs in both English and Afrikaans versions, and is of a multiple choice, pencil-and-paper format. It is suitable for testees who have had 12 years of education or more (Standard 10 or higher). The battery consists of six subtests, which together require approximately three hours to administer and complete (NIPR, 1975). The manual provides norm tables obtained from groups of White matriculants, first year university students, college students, Black bursary applicants, Indian first year university students and White university graduates. Subtests for this battery are as follows:

(a) Mental Alertness (45 minutes)

This is a test of general intelligence comprising 42 items. These include numerical and letter series, verbal analogies, common elements and other problems requiring reasoning ability.

(b) Arithmetical Problems (40 minutes)

This test has 20 verbally-stated arithmetical problems,requiring testees to solve these.

(c) Reading Comprehension (20 minutes)

A test of language ability on the basis of four short paragraphs. Testees are required to answer five questions based on the content of each of the four paragraphs.

(d) Vocabulary (12 minutes)

In this subtest subjects are presented with 40 English words, each followed by 5 words, only one of which is synonymous with the given word. Testees have to choose the right one.

(e) Lees- en Begripstoets (20 minutes)

An Afrikaans Reading Comprehension test parallel to the English version described above.

(f) Woordeskattoets (12 minutes)

An Afrikaans Vocabulary test parallel to the English one described above.

2.4 THE INTERMEDIATE BATTERY AND THE IMAT

2.4.1 The Intermediate Battery

As mentioned in Section 2.1, the Intermediate Battery was developed "de-novo" by the NIPR in response to the need of the South African Railways for an instrument to select personnel with a "medium" level of education. As with the other batteries, the Intermediate Battery has English and Afrikaans versions, and is of a multiple choice, pencil-and-paper format. It has been found suitable for people with 9-12 years of education (Standards 7 to 10). The battery consists of 7 subtests, together requiring three and a half hours to administer and complete (Wilcocks, 1973). The manual(Wilcocks,1973) provides norm tables obtained from groups of White male clerical workers and college students. Subtests for this battery are as follows:

(a) Mental Alertness (30 minutes)

This will be covered in detail below.

(b) Arithmetical Problems (45 minutes)

Containing 20 verbally-stated arithmetical problems,this test requires testees to solve these.

(c) Computation (35 minutes)

This test has 30 items designed to check computational ability. As such it has problems of addition, subtraction, multiplication, division, fractions, decimals etc.

(d) Spot-the-Error (10 minutes)

Comprising 60 items, this test was designed to check certain clerical abilities. Testees are required to find errors in a list of names, titles, weights and measures by comparing them with a "correct" list.

(e) Reading Comprehension (20 minutes)

Once again a test of language ability on the basis of four short paragraphs. Subjects are required to answer five questions based on the content of each of the four paragraphs.

(f) Vocabulary (10 minutes)

In this test subjects are presented with 45 items for which they must choose the correct respective meanings.

(g) Spelling (15 minutes)

In this test 30 words are each spelt in 5 different ways. From the alternatives, subjects are required to choose the correct spelling for each word.

2.4.2 **Description of the IMAT**

The IMAT is one of seven subtests of the Intermediate Battery. The IMAT may be regarded as a measure of verbal problem solving ability,requiring the ability to apply previously learned knowledge to the solution of problems. The IMAT has 30 items, which are similar to those commonly found in tests of general intelligence. They include codes, similarities, number and letter series, and analogies (Wilcocks, 1973).

As with other subtests of the Intermediate Battery, the IMAT is contained within a re-usable booklet, and has a multiple choice format. The test is scored by hand with the aid of a stencil. This differentiates the IMAT from the Mental Alertness (Intermediate), which requires the testee to write down the correct answer in the test booklet. Furthermore, some of the items of the Mental Alertness (Intermediate) are not of a multiple choice format and answers have to be generated by the testee.

The Intermediate Battery was standardised on White male clerical staff in the main centres of the South African Railways between 1964 and 1965. It was also standardised on a relatively smaller sample of White male and female student teachers at the same time. Details of these samples are included under the list of available norms in Appendix 1.3.

2.4.3 **Revision of the IMAT**

It has been mentioned above that with the metrication of tests in 1973, the A form of the IMAT was discarded since it would have required extensive item revision. Conversely the B form was maintained,since it required little modification of items. The current edition of the IMAT differs from the original B form only on item numbers 4 and 24. Originally item 4 referred to ounces, now it refers to kilograms. Item 24 originally referred to inches and feet, and now it refers to millimetres and centimetres.

2.4.4 **Computerisation and the IMAT**

Computerisation of tests offers the prospect of a number of general advantages over conventionally administered tests (Taylor, Werbeloff and Ebertsohn, 1982). These include:

- standardised administration
- instructions match the specific pace of the testee
- testing can be conducted at more or less whatever time is convenient for the testee
- persons with minimal psychological training are able to administer the tests
- scoring and norming is instantaneous and error free
- confidentiality and record keeping are more easily managed

Bearing in mind these kinds of advantages, in 1978 the Test Construction Division of the NIPR began devoting time to the computerisation of tests with the ultimate aim of offering a spectrum of computerised tests for selection and placement purposes. Work was initially focussed on programming the NIPR tests, including the IMAT, on a mainframe system known as PLATO. Since there are a number of significant administrative differences between a pencil-and-paper test and its computerised counterpart (Taylor,Werbeloff and Ebersohn, 1982), part of the researchers' work involved validating the newly computerised tests against the originals. A validation study was carried out specifically on the IMAT by Taylor et al (1982) which is discussed in detail in Chapter 6. The results showed that the mean difference between paper-and-pencil and computerised versions of the IMAT for both English and Afrikaans research samples were low and insignificant. Both English and Afrikaans versions of the computerised IMAT were found to have acceptable Kuder-Richardson reliability values and the results of their intercorrelations with other tests (see Chapter 6) led the researchers to conclude that there were no significant differences compared with the performances of the pencil-and-paper versions. Subsequently other computerised versions of the IMAT have been developed for personal computers following the trend away from mainframe computers to personal computers.

CHAPTER 3

RELIABILITY AND VALIDITY

In the context of a given industrial, educational or clinical setting, a test user faces the problem of having to choose the most appropriate of a variety of possible tests. What is required is some way of comparing tests in terms of performance and suitability. The concepts of validity and reliability can provide valuable information aiding in the choice and evaluation of tests in given contexts. These concepts will be focussed on in this chapter. However, bearing in mind the central role of correlation to the concepts of reliability and validity (for example, see Anastasi, 1982) it may be useful to begin by highlighting some important issues in the interpretation of correlational data.

3.1 CORRELATION

It is not the intention to provide a detailed discussion on the different types of correlation and their respective computations, but rather to provide information which will clarify the use and interpretation of correlation. A more technical discussion on correlation may be found in most books of psychological statistics such as Guilford (1965), Guilford and Fruchter, (1978), Huysamen (1981) etc.

3.1.1 Statistical Significance, and Type I and Type II Errors

Correlation values obtained from small samples are likely to have larger standard errors, and sample sizes vary widely from study. It is important to be able to generalise the results from a particular sample of individuals to the wider population which the sample represents. The concept of statistical significance has been developed as a tool, which in the instance of correlation, predicts likelihood of a given correlation coefficient calculated for a given sample being representative of the wider population (Huysamen, 1981).

Statistical significance of correlations is derived by assuming that the correlation of the variable under investigation in the wider population is zero, and then establishing the likelihood of a correlation as high as the one observed in the sample being attributable to chance factors (Anastasi, 1982). Statistical significance is usually presented at either the 5 percent or the 1 percent level. If a correlation is found to be significant at the 5 percent level it could be concluded that it is highly probable that the correlation for a given sample is representative of a correlation in the wider population (i.e. there are only 5 chances in a hundred of being wrong in this conclusion). As the significance level increases (e.g. 1 percent or 0,1 percent),so the likelihood increases that the correlation for a given sample is representative of a correlation in the wider population (Anastasi, 1982).

The significance test, then, has been derived in response to the problem of sampling error. Based on what has been said above,it may be assumed that the use of significance tests guarantee an error rate of 5 percent or less. However, the 5 percent error rate (or less) may only be guaranteed if the null hypothesis for a given study is true, i.e.. that the correlation of the variables in the population is zero. If the null hypothesis is false, the error rate may go up to 95 percent (Guilford, 1965). These premises underlie what is known as Type I and Type II errors. Type I errors occur if the null hypothesis is true for the population and given sample data lead one to reject it. Type II errors occur if the null hypothesis is false for the population and given sample data lead one to accept it (Guilford, 1965).

In interpreting data from a given study it is important to know whether one is dealing with a 5 percent or 95 percent error rate. If it is known that the null hypothesis is true, then it can be concluded that the significance test has an error rate of 5 percent - but in this case the data could be ignored (Hunter, Schmidt and Jackson, 1982). If it is not known whether the null hypothesis is true or false, then it is not known whether the error rate is Type I or Type II, and one faces high risks in the application of findings to the population from which the sample was drawn. To illustrate this point, Hunter, Schmidt and Jackson (1982) present a table of the correlational data on the relationship between organisational commitment and job satisfaction of thirty different studies. Using standard review practices (such as attempting to find patterns in the correlational data according to sample variables such as sex, race, age, job level and geographical location) the authors come up with a number of plausible hypotheses to explain disparate findings that are observed across the studies,before revealing that the table in fact represents thirty variations of a single correlation of 0,33, and that the observed variation is solely attributable to the effect of sample size (sample size was randomly chosen from a distribution centring around 40). Hunter et al (1982) proceed to show that 31 percent of findings across the thirty studies were affected by Type I or Type II errors, and comment that error rates of over 50 percent have been shown to be the usual case in personnel selection research - resulting in "conflicting results" across studies.

From what has been said above, it can be seen that the uncritical use of significance tests can lead to erroneous conclusions in review studies, and therefore an alternative method of interpretation is indicated. There are two alternatives to the significance test.When considering a body of related studies meta-analysis can be used (see Section 3.3.7), and when considering a single study, confidence intervals can be used.

3.1.2 Standard Error and the Confidence Interval

Faced with a given correlation coefficient, such as correlating performance on the IMAT with a certain criterion of job success, it is useful to establish how representative the correlation is of the "true" correlation between test and criterion (i.e. the mean of the distribution of correlations that would occur if the correlation were hypothetically carried out an infinite number of times - Guilford, 1965).

Clearly, it is not possible to carry out a given correlational study an infinite number of times in order to establish the "true" correlation. However, bearing in mind the standard properties of a standard deviation unit, standard error enables an estimate to be made of the likelihood of a given observed correlation approximating the "true" correlation concerned. This may be achieved by establishing the confidence interval of a given observed correlation, which in effect is the estimated variance range between which the observed correlation is expected to occur (Hunter et al, 1982). Calculation of the confidence interval may be found in most psychology statistics textbooks such as Guilford (1965). Confidence intervals for given observed correlations are provided in statistical tables such as Neave (1978), allowing for ease of reference. Confidence intervals are usually reported at one of two levels of confidence, either the 5 percent level or the 1 percent level. At the 5 percent level there is a deviation that leaves 5 percent of the area in the two tails of the normal distribution, and at the 1 percent level there is a deviation which leaves 1 percent of the area in the two tails of the normal distribution.

Hunter et al (1982) argue that confidence intervals are superior in the interpretation of given observed correlation values for two reasons. Firstly, the interval centres on the observed value rather than on the hypothetical value of the null hypothesis, and secondly it provides a clear indication of the extent of uncertainty in small sample studies (which is often overlooked when dealing with a single figure). As is covered in Sections 3.3.6 and 3.3.7, Schmidt, Hunter and their colleagues argue that the only way to eliminate uncertainty is to either run large-sample, single studies, or to combine results across many small-sample studies through the use of meta-analysis.

3.1.2 **Limitations of Correlational Studies**

The relative simplicity of the correlation coefficient, both to compute and to understand, make it a popular statistic in research (Anastasi, 1982). While recognising that well conducted studies utilising correlations can provide valuable sources of data, there is strong need for caution when interpreting correlational data. In addition to the problems of interpretation of significance mentioned above, there are a number of reasons:

- Correlations coefficients are affected by the range of individual differences within a group. Experimental samples that are highly homogeneous in relation to attributes relevant to the correlation measurements e.g. education, or given abilities, are likely to yield lower correlations than samples where such attributes are more heterogeneous (Anastasi, 1982)

- Research conditions are never exactly repeated. There are a wide variety of possible variables that can differ across different settings, and which can affect the relationship between two given variables. Such variables include the characteristics of the experimental sample, experimental method, equipment used, methods of training or instructing subjects, motivation or attitudes of experimenters or subjects, and the reliability and validity of the measures used to evaluate the criteria (Guion, 1980). Such variables form the central issue of contention in the concept of validity generalisation covered in Section 3.3.6

- Correlation coefficients are affected by any contaminates that affect either or both the two variables being related (Tukey, 1969).

Therefore, when correlation data is being evaluated, the following considerations should be checked:

- how appropriate is the sample in terms of size, range, and representativeness of the population being represented?
- how homogeneous/heterogeneous is the sample in relation to attributes relevant to the correlation measurements eg. education, age?
- is the research design logical and consistent? Have standardised procedures been followed to ensure replicability of results (eg. Selltiz, Jahoda, Deutsch and Cook, 1959).
- have the measuring instruments used to evaluate the criteria been assessed for validity and reliability? If so, is the validity and reliability acceptable?

3.2 RELIABILITY

For a test to be useful it must measure consistently, otherwise little weight could be placed on its results, and it would be unethical to base any important decisions on such results. The main threat to the consistency of a test comes from the interference of variables irrelevant to the purpose of the test, which may otherwise be called sources of error variance. In taking measurements in other sciences, physical properties such as length, weight and temperature are more condusive to consistent measurement, since it is relatively easy to control for sources of variance when dealing with such properties (Thorndike and Hagen, 1969). Conversely psychological or behavioural properties are relatively harder to measure consistently since sources of variance affecting the testing process are more difficult to control. Examples of such variables include whether the examinee is feeling physically and emotionally well, whether the examinee has had any recent exposure to information better enabling him to answer the test items, and whether there are any factors negatively affecting the test administration such as background noise (Aiken, 1982).

Applied to testing the concept of reliability refers to the consistency with which a test measures over different administrations involving different occasions, test forms, scorers etc. (Huysamen, 1988). Underlying the calculation of reliability is the rationale that any observed test score comprises a "true score" (directly reflecting the attribute being measured) and "error score" (reflecting sources of error variance). 'Put another way, it assumed that differences that are observed across different test score administrations are attributable to:

a. Genuine differences in the attribute being measured ("true score" variance) eg. mechanical comprehension before and after a 3 year course in mechanics

b. Contaminating variables which are irrelevant to the purpose of the test (error variance)

Test reliability measurements help to indicate the extent to which differences in test scores are attributable to true differences in what is being measured, and the extent to which these are attributable to chance errors (Anastasi, 1982).

There are various measures of test reliability which are sensitive to different sources of error variance, and this makes it possible to estimate what proportion of observed score variance can be attributed to true score variance, and what can be attributed to error score variance. These will be outlined below, but firstly it will be useful to differentiate the different sources of error variance.

Broadly, there are two main sources of error variance, namely temporal and item sampling sources (Anastasi, 1982). Temporal sources of error variance are contaminating variables linked to taking tests on different occasions. These can be divided into administrative variables, external variables and individual variables. Administrative variance arises specifically from poor administration or scoring of a test, which could include inconsistent instruction giving or timing, careless marking, inconsistent test interpretation (for example in interpreting personality tests), and poor test control (resulting in testees obtaining prior exposure to test material). Clearly poor test administration will negatively affect test reliability, and one will therefore attempt to control for such sources of variance through proper control of tests, and proper training of test users (see Prinsloo,1982).

Sources of external variance include contaminates such as noise distractions and extreme temperature or lighting fluctuation (Aiken, 1982). Sources of individual variance include contaminates such as illness, fatigue, emotional strain, recent experiences leading to distracting thoughts or lucky guessing.

Concerning the second main type of error variance, item sampling sources of variance are contaminating variables which arise from a particular selection of examination items. Thus a testee may unduly benefit or be disadvantaged by the degree to which a test covers specifically what he or she has learned for, or been fortuitously exposed to (Thorndike and Hagen, 1969). The different measures of test reliability shall now be covered.

3.2.1 **Test - Retest Reliability**

The test-retest method of establishing reliability of test scores simply requires correlating the scores obtained by given individuals from two administrations of the same test. This type of reliability coefficient shows the extend to which scores on a test are stable over time, and are resilient to random changes in the condition of the examinee or testing environment (i.e. temporal sources of error variance).

In establishing test-retest reliability one of the crucial factors is that of the time lapse between the two test administrations (Anastasi, 1982). If the time lapse is too short, examinees may benefit from a memory effect enabling them to recall their previous responses (and thus yielding an underestimate of the retest reliability). On the other hand if the time lapse is too long, there may be significant changes in the attribute being measured (this is particularly pronounced in children, whose personalities, aptitudes and interests develop rapidly). Thus, where test-retest reliability is reported it needs to be accompanied by a statement of the length between test administrations and any relevant intervening experiences affecting testees e.g. training (Anastasi, 1982).

3.2.2 **Alternate Form Reliability**

Alternate form reliability is calculated by correlating the scores obtained by given individuals on parallel forms of the same test. The same individuals can thus be tested with one form on the first occasion and the other, comparable form, on the second. This type of reliability coefficient is primarily sensitive to sources of error variance arising from item sampling (Huysamen, 1980).

Bearing in mind that the use of alternate test forms minimises the memory effect, it is possible to retest soon after the administration of the first form. In fact failure to do this risks the influence of sources of error variation arising not only from item sampling, but also from temporal variables (Anastasi, 1982). Bearing this in mind, reports of alternate form reliability should again be accompanied by information regarding the time lapse between testing, as well as any relevant intervening experiences.

3.2.3. **Measures of Item Homogeneity/Internal Consistency**

Test homogeneity refers to the extent to which the different items in a test are consistent with each other in terms of what they are measuring. While there is some debate as to whether item homogeneity is really an aspect of reliability (Anastasi, 1982), the homogeneity of a test will affect the consistency of interpretation of different scores on that test. Interpretation of a test with homogeneous items is likely to be less ambiguous than one with heterogeneous items (Anastasi, 1982). For this reason, in instances where a test is intended to measure a complex phenomenon (e.g. general aptitude), it may be more desirable to construct several relatively homogeneous sub-tests than one heterogeneous test. There are three ways of measuring item homogeneity, namely: split-half reliability, Cronbach's coefficient alpha, and Kuder and Richardson's reliability. Each of these measures is sensitive to item-sampling sources of error variance.

3.2.3.1 Split-Half Reliability

The principle underlying this form of reliability is that a single test is divided into two parts such that these may be regarded as two parallel tests (Huysamen, 1980). Performance on these two parts are then correlated with a view to providing a coefficient of internal consistency. There are a wide range of ways that a test may be divided into two halves, and correlation coefficients can vary from one split to the next. Caution must therefore be exercised to ensure that the two halves are comparable. In some cases this can be achieved by dividing the test into odd and even items.

Once the two test halves have been established, and individual performances correlated, it is important to correct for the fact that the reliability coefficient has in effect only been derived from half a test. (Particularly since the reliability of a test increases with length - Nunnally, 1967). This can be achieved by means of the Spearman-Brown formula (which is also useful for determining the effect on reliability of increasing the length of any test), which is as follows (simplified after Huysamen,1988):

$$reliability\ of\ lengthened\ test = \frac{J \times reliability\ of\ original\ test}{1 + (J - 1) \times reliability\ of\ original\ test}$$

where J = the number of times the test has been lengthened

3.2.3.2 Cronbach's Coefficient Alpha

Bearing in mind that a split-half reliability can vary from one split to the next, and that it can be difficult to ensure that the halves selected are indeed comparable, a reliability coefficient has been devised which represents the mean of all possible split-half coefficients. (This provides a better estimate of test reliability than any single split alone). This is known as Cronbach's Coefficient Alpha (Cronbach, 1951) which can be calculated as follows (simplified after Huysamen,1988):

$$coefficient\ alpha = \left\{\frac{J}{J-1}\right\}\left\{1 - \frac{sum\ of\ the\ item\ variances}{variance\ of\ the\ total\ test}\right\}$$

where J = the total number of items in the test

Similar to alternate form reliability and split-half reliability, coefficient alpha is affected by error variance arising from content sampling (Anastasi, 1982). However, it is also affected by the degree of homogeneity of areas covered by test items (e.g. a single test with items covering different skills such as vocabulary, mathematics and spatial ability would tend to show less inter-item consistency than one with only vocabulary items).

Coefficient alpha is applicable to tests which have multiple scored items (e.g. on a personality test which may have four score options, each of which are assigned a different score value - Cronbach, 1951).

3.2.3.3 Kuder and Richardson Reliability

This reliability index has the same rationale as the coefficient alpha, and was developed by Kuder and Richardson (1937). Unlike the coefficient alpha, the Kuder and Richardson formula is applicable to dichotomously scored items (e.g. right or wrong). In the original article, two important formulae were derived, respectively Kuder and Richardson Formula 20 (KR-20) and Kuder and Richardson Formula 21 (KR-21). The former can be shown to be algebraically equivalent to coefficient alpha (Huysamen, 1980). KR-21 is derived under the assumption that all the items in the test are equally difficult. If this assumption is false KR-21 will be smaller than KR-20. In practice it is therefore desirable and simple to compute KR-20 and KR-21. However, if KR-21 is computed in isolation and found to be satisfactorily high, it is unnecessary to calculate KR-20 since this will be at least as high as the value obtained for KR-21 (Huysamen, 1988). This is useful to know, bearing in mind that the KR-20 formula is more cumbersome to calculate,requiring an item analysis(Kuder and Richardson,1937).

The formula for KR 20 and KR 21 are presented below (after Kuder and Richardson, 1937):

$$KR\text{-}20 = \left[\frac{K}{K-1}\right]\left[\frac{St^2 - Xt + \frac{Xt^2}{K} + KSp^2}{St^2}\right]$$

$$KR\text{-}21 = \left[\frac{K}{K-1}\right]\left[\frac{St^2 - Xt + \frac{Xt^2}{K}}{St^2}\right]$$

Where:

K = Number of items in the test

Xt = Mean of the test score

St = Standard deviation of the test score

Sp = Standard deviation of items pi (where candidates give the correct answer to item i)

3.2.3.4 The Tucker Correction and Kuder and Richardson Reliability

Because it is calculated relatively simply, the KR-21 formula is often used to calculate reliability. However, this formula sometimes seriously underestimates the reliability of a test (Tucker, 1949). The KR-20 formula yields a much better estimate, but as noted above requires results from an item analysis of the test,which is relatively cumbersome to perform. Therefore a correction factor has been devised for use with the KR-21 formula (Tucker, 1949). This correction factor is known as the Tucker correction. The Tucker correction is equivalent to the K Sp2 value in the KR-20 formula. It can be seen from the formula below that the correction is inversely proportional to the raw score standard deviation of the test. This means that the lower the standard deviation value the higher the correction value, and vice-versa (Tucker, 1949).

$$KR\text{-}21 = \left[\frac{K}{K-1}\right]\left[\frac{St^2 - Xt + \frac{Xt^2}{K}}{St^2}\right] + \frac{0{,}03K^2}{(K-1)\ St^2}$$

KR-21 + Tucker Correction (simplified after Wilcocks, 1973)

Some of the main types of reliability which cover different types of error variance have now been considered. Two instances of reliability which have less relevance to this study, but which may be noted, concern speeded tests and tests which leave a good deal of judgment to the scorer. In the former instance, single-trial reliability coefficients such as found by odd-even or Kuder and Richardson techniques, are inapplicable. In such instances, test-retest or equivalent form techniques are more appropriate (Anastasi, 1982). Concerning tests requiring a good deal of scorer interpretation (such as personality tests), one of the ways of ascertaining reliability is by having a sample of answer sheets independently scored by two examiners whose results are then compared.

3.2.4 **Interpreting Reliability**

3.2.4.1 Magnitude of the Reliability Coefficient

Since reliability coefficients are in fact correlations, the closer that they are to 1.00 (representing a perfect, positive relationship) the better, for a test cannot be too reliable. While some frequently used tests have test- retest reliabilities only in the 0,50 range, tests with coefficients in the 0.70 range are usually professionally acceptable (Muchinsky, 1986). An important proviso for evaluating test reliability is how the test scores will be used. If a test is to be used for group prediction, a reliability of 0.65 may be considered satisfactory. However, a test used for individual prediction (for example to diagnose brain impairment), should have a reliability of at least 0.9 (van den Berg, 1989). It is important to note that before drawing conclusions about the reliability of a test, the obtained correlation should be checked to see that it is statistically significant at an acceptable level, or that the possible variance range is taken into account (see Section 3.1.1 and 3.1.2)

3.2.4.2 Factors Affecting Reliability Coefficients

Earlier in this section it was stated that the rationale underlying the calculation of reliability is that any observed score comprises a "true score" and an "error score". Based on this premise it is possible to define reliability as the ratio of true score variance to observed score variance (Huysamen, 1980). From this it can be shown mathematically that whenever observed variance is increased and error variance remains fixed, the reliability figure will become greater. (Aiken, 1982). Several ways have been found to increase observed variance without appreciably increasing error variance (thus increasing stated reliability), including:

-increasing the length of the test (cf the Spearman - Brown formula)

-composing a test with items of moderate difficultly (which has been shown to lead to greater variance than composing a test with very difficult or very easy items - Aiken, 1982).

Bearing the above in mind it is important that every reliability coefficient is accompanied by a full description of the type of group on which it was determined. It is important to note that a reported reliability coefficient is applicable only to samples similar to the one which it was computed (Anastasi, 1982)

3.2.4.3 Reliability and the Standard Error of Measurement (SEM)

Implicit in the notion that a given observed score is made up of a "true score" component and an "error score" component, is that a given score obtained by a testee will fall on a normal distribution of possible scores due to the sources of error variance mentioned above. Theoretically an individual "true score" would represent the mean of the normal distribution of possible scores for a given testee. The SEM is in effect the standard deviation of the possible score distribution for a given testee on a particular test. Bearing in mind the standard properties of a standard deviation unit, SEM enables an estimate to be made of the likelihood of a given observed score approximating an individuals "true score" (Anastasi, 1982). SEM is calculated as follows (simplified after Huysamen,1988):

$$SEM = \text{Test score standard deviation} \times \sqrt{1.00 - \text{reliability}}$$

3.3 VALIDITY

Applied to testing, the concept of validity is concerned with whether a test measures what it is supposed to measure, and how well it does so. The process of validation cannot be done in the abstract, but must rather be considered in the context in which the tests will be used (Anastasi, 1982). Moreover, validation is a matter of degree rather than an absolute property, and it takes place on an ongoing basis (Nunnally, 1967). All procedures for determining test validity are concerned with the relationship between how a test performs and some criterion of the attribute which they are intended to measure. If a test is intended for more than one purpose it will have a separate validity for each purpose, for a test that is valid for one purpose may not be valid for another (Huysamen, 1988).

There are several different types of reliability, and these will now be briefly covered.

3.3.1 Face Validity

Face validity is the simplest understood of the different types of validity. In fact technically it is not "validity" at all, in that it does not refer to what a test actually measures, but rather what it **appears** to measure. Face validity is an important feature of a test because if it appears irrelevant or invalid it may not be accepted as useful or legitimate by the intended target group of the test. However, face validity should not be considered a substitute for objectively determined validity such as is outlined below.

3.3.2 Criterion-related Validity

Although all forms of validity are evaluated by means of some form of criterion, traditionally the term "criterion-related" validity has been used to refer to predictive or concurrent validity (Guion,1980).In the case of predictive validity the criterion against which a test is evaluated is some future performance or status, such as academic achievement or on-the-job performance. While the predictive validity of a test is evaluated against some future criterion or predictor, concurrent validity is assessed against an existing criterion, such as the score on another test, or existing performance records, or a contrasted group (Muchinsky, 1986). The primary difference between predictive and concurrent validity is that of the time lag between testing and the availability of the criterion measure. Concurrent validity may be assessed if it is impractical to extend validation procedures over time, or in instances where tests are being used to diagnose existing status rather than prediction of future outcome (Anastasi, 1982). For example concurrent validity may be evaluated when ones test is being proposed as a substitute for another (assuming it has some advantage over the other), or in seeing if a test is able to differentiate between types of psychiatric disorder.

It may be noted that the relationship between the predictor and the criterion is of primary concern in criterion - related validation. While it may be desirable, there is in fact no need to know what specific psychological phenomena are accounting for the test performance (Meehl, 1973).

3.3.3 Content Validity

Content validity refers to the extent to which a test covers a representative sample of the behaviours or phenomena which it is supposed to measure. The concept originated from educational measurement (Guion, 1980), where a given test could be considered a valid measure of curriculum content insofar as it adequately represented course material. Clearly content validation is a process which must begin when a test is being constructed. This is usually a deductive process, ideally conducted along the following lines (Helmstadter, 1966):

- Defining the universe under consideration
- Dividing the universe into constituent categories
- Constructing and sampling items from each category

Guion (1980) notes that though ultimately worthwhile, these steps are laborious and that content validity ultimately requires rigour in each. He particularly warns of the difficulty of devising an effective scoring system that minimises the influence of contaminates such as anxiety, inability to comprehend verbal instructions, or particular perceptual skills that favour one type of testee unfairly over another.

Once constructed, the content validity of a given test can be assessed by procedures such as: analysing the types of errors commonly made on the test, getting testees to think aloud while performing the test, comparing performance on a test before and after a period of relevant training, and correlating it with tests purporting to measure the same thing (Lawshe, 1975, Aiken, 1982, Nunnally, 1967).

3.3.4 Construct Validity and the Nomological Network

A construct is a postulated attribute of people, assumed to be reflected in the test performance (Meehl, 1973). Examples include personality, motivation, interest, intelligence etc. Put simply, validity is the extent to which a test may be said to measure a theoretical trait or attribute (Huysamen, 1988).

It was mentioned above that all forms of validity are evaluated against some criterion.Face validity is assessed against whether a test looks as if it is measuring what it is supposed to be measuring. In concurrent or predictive validity, the relationship between the test and some existing or future criterion is of primary concern (note that these forms of validity have little to do with what the test is specifically measuring, but rather the extent of their relationship with the criterion). Content validity is evaluated by assessing the representativeness of the test content against the subject universe which the test is sampling. In contrast, because constructs are relatively broad and abstract concepts, researchers have no definite criteria against which to measure the construct validity of a given test, and have to gather and integrate information from a number of "partial" criteria (Meehl, 1973). In fact, it is possible that initially at least, researchers have little more than a vague conceptual definition of the construct they wish to measure, (Guion, 1980). However, through a systematic process of piecing information together from a variety of sources, it is possible to obtain a rich knowledge of a given construct. There are a wide variety of sources of evidence for construct validity (Meehl, 1973; Anastasi, 1982; Nunnally, 1967; Guilford, 1948; Eysenck, 1980; Cronbach, 1949), which include:

1) Use of the technique of factor analysis. For example, if a number of tests are used on a test sample, factor analysis of test performances can reveal correlation clusters that suggest the location of common constructs.

2) Linked to 1), positive correlations with other tests purporting to measure the same construct.

3) Differences in test performance of contrasted groups (for example education differences) on a test of general aptitude (a construct which is known to be affected by education - Owen, 1989).

4) Age differentiation in performance on a test. For example the traditional notion of "intelligence" assumes that abilities develop with age, at least until maturity (of the age norms of the Stanford-Binet test).

5) Analysis of the internal consistency of the test. If the items of a test are truly measuring a given construct, one would expect there to be a fair degree of intercorrelation between them. Meehl (1973) comments that negative item - test correlations may support construct validity if the items with negative correlations are held to be irrelevant to the postulated construct.

6) The stability of test scores in a test-retest design, particularly after some controlled experimental intervention not affecting the construct which the test is supposed to measure.

7) Observation of a testee's process of performance. An individual's approach to a task or test item may give valuable insights into the postulated construct, as well as whether the task or item is being approached in the way it was designed to be. Meehl (1973) gives the example that observation of testee performance on a particular test showed that erroneous reading of the question was common, influencing the interpretation of low scores.

(8) Convergent and discriminant validation. The construct validation of a test can be demonstrated not only if it correlates highly with other tests of the same characteristic (convergent validation), but also if it has low correlations with measures of different characteristics (discriminant validation). Campbell and Fiske (1959) propose a systematic experimental design for the evaluation of convergent and discriminant validity which they called the multitrait-multimethod matrix. This involves correlation between measures of:

(a) the same trait by the same method
(b) different traits by the same method
(c) the same traits by different methods
(d) different traits by different methods

Evidence for construct validity occurs when one finds that correlations between the same trait measured by the same and different measures are higher than the correlations between different traits measured by the same or different measures.

As implied above, constructs are inferred theoretical postulates.One of the difficulties posed in attempting to establish construct validity is that by virtue of their nature, constructs cannot be operationally defined.The logic of construct validation therefore requires that for a construct to be scientifically supported,it should occur within a nomological network. A nomological network is defined as "an interlocking system of the laws which constitute a theory" (Meehl, 1973, p.16). In effect, a nomological network is a systematic way of bringing together evidence for a given construct,such as those listed in the points 1) to 8) above, which provides logical evidence for the existence of that construct.

Meehl (1973) describes the framework of nomological networks as follows:

1) In order to make the existence of psychological constructs clear in a way which is scientifically acceptable, one needs to establish the laws in which they occurs.

2) The laws in a nomological network may be statistical or deterministic, and may relate observable properties or quantities to one another; theoretical constructs to observables; or different theoretical constructs to one another.

3) A necessary condition for a construct to be scientifically admissible is that it occurs in a nomological network, at least some of whose laws involve observables.

4) Learning more about a theoretical construct is a matter of elaborating the nomological network in which it occurs, or increasing the definiteness of its components.

5) Enrichment of the network such as adding a construct or a relation to theory is justified if it generates "nomologicals" that are confirmed by observation, or if it reduces the number of "nomologicals" required to state a given assertion.

6) Unless a given network makes contact with observations and shows clearly explicit steps of inference, construct validation cannot be truly claimed.

3.5 Interpreting Validity

The different "types" of validity create the potential for considerable confusion regarding how they relate to one another. Guion (1980) comments that some people seem to regard these as separate entities, and points out that content, criterion - related and construct validity are in fact different manifestations of one concept. As such the different validity types serve as evidence for the single attribute of validity. As a relatively complex concept, it is important that validity is not confused with a given statistical score. In fact validity values have no absolute meanings, but rather enable one to draw various inferences about validity (Guion, 1980).

Validity, then is a property of inferences of scores. These scores reflect the way that a given definition of validity has been operationalised. There are a variety of areas for methodological sources of error to affect validity scores, ranging from the adequacy of the sample size, to the logical foundation of the operational hypothesis, to the quality of the measures of the variable (Guion, 1976). While the simplicity of the validity coefficient makes it an attractive measure for

the study of validity (and reliability), it is subject to a number of limitations (covered in section 3.1.3 above). It follows that reliance on a single validity value or even a given study when interpreting validity data should be avoided. Validity data are much more powerful when interpreted within the context of other related studies (cf the rationale of the nomological network).

3.3.5.1 Magnitude of the validity coefficient

It is difficult to provide a generic guideline as to how high a validity coefficient should be, since such a coefficient should be interpreted in its context. However, it can be borne in mind that just as a test cannot be too reliable, it cannot be too valid, and therefore the closer the coefficient is to 1.00 the better (Muchinsky, 1986). Furthermore, as with the reliability coefficient, before interpretations are made concerning validity, the obtained coefficient must be high enough to be significant at an acceptable level,or preferably, interpreted within the parameters of confidence intervals (see Section 3.1.2). Validity coefficients are usually interpreted in one of three contexts, namely:

a) if one wishes to determine which of two tests is more effective in predicting the exact criterion scores of individuals (in which case the standard error of estimate is considered)

b) if a particular test has been chosen and one wishes to assess the extent to which it will result in errors of prediction (in which case the standard error of estimate is considered).

c) if a particular test is to be used as a selection instrument (in which case its success ratio is considered)

Each of these instances will now be discussed, and it will be seen that in certain instances tests with validities as low as 0.20 or 0.30 can be useful (such as when being used as part of a selection program). In other instances tests with validities as high as 0.80 may have limited usefulness (for example if a test was required to predict an individual's exact position in the criterion distribution - Anastasi, 1982).

3.3.5.2 The Coefficient of Determination (COD)

Where two tests are being compared on their ability to predict exact criterion scores, the COD is a useful indicator. The COD is the square of the validity coefficient .Just as the reliability coefficient indicates the proportion of observed score variance attributable to true score variance, the COD reflects the proportion of criterion score variance attributable to test score variance (Huysamen, 1988). The COD needs to be quite high before a sizable proportion of criterion score variance is explained by test score variance rather than other, unknown factors

3.3.5.3 The Standard Error of Estimate (SEE)

This index may be used when one wishes to assess the extent to which a particular test may make errors of prediction. SEE is analogous to the standard error of measurement discussed in connection with reliability (see section 6.5.6.3). Just as the standard error of measurement indicates the margin of error to be expected in an individual's score as a result of the unreliability of a test, the SEE shows the margin of error to be expected in an individuals predicted criterion score as result of the imperfect validity of the test. SEE is calculated as follows (simplified after Huysamen,1988):

$$SEE = \text{criterion score standard deviation} \times \sqrt{1.00 - \text{squared validity coefficient}}$$

The higher the SEE, the greater the deviation of the predicted criterion scores tends to be from the actual criterion scores.

3.2.5.4. The Success Ratio

The success ratio is an index that is useful to consider when a test is being specifically used as a predictor(Anastasi, 1982). For example a given test is being evaluated for its ability to predict job success. Firstly a test cutoff score needs to be established, above which job applicants will be accepted, and below which job applicants will be rejected. Secondly, it is necessary to determine some measure of job success against which the performance on the test can be evaluated. This makes it possible to test a sample of job applicants, and to correlate their test performances against the measure of job success. Hopefully the majority of those who attain the test cutoff score or above will also meet the job criterion. These are known as valid acceptances. Those obtaining the test cutoff but failing the job criterion are known as false acceptances; those failing both test cutoff and job criterion are known as valid rejections, and those failing the test cutoff but passing the job criterion are known as false rejections. The success ratio refers to the proportion of predicted successes who in fact turn out to be successful i.e:

$$\text{Success ratio} = \frac{\text{number of applicants who pass the test cutoff}}{\text{number of the above who pass the job criterion}}$$

It is clear that the higher the validity coefficient of the test, the higher the success ratio. However the success ratio does not depend only on the validity of the test but also the selection ratio, or the ratio of the number of vacancies to the number of applicants. It can be shown that if the selection ratio is low (i.e there are few vacancies relative to the number of applicants), then a test with even a low test - criterion ratio can be very useful (Taylor and Russell, 1939). To aid in making decisions pertaining to the use of predictors in selection, Taylor and Russell (1939) developed a series of tables. Dimensions of the tables include:

a) the base rate (i.e proportion of successful applicants selected without the use of a predictive test)
b) validity coefficient of the predictor
c) the selection ratio (i.e the ratio of the number of vacancies to the number of applicants).

The tables indicate the number of "successes" to be expected through the use of a test of given validity, selection ratio and base rate. For example, assuming that a testing programme is to be introduced to a organisation that has previously selected on an ad-hoc basis, it is firstly necessary to establish the success rate of applicants selected under the ad-hoc system. With this as a "norm" the Taylor-Russell tables can be used to estimate the test validity and selection ratio necessary to make a significant improvement on the ad-hoc selection method.

3.3.6 **Validity Generalisation**

Criterion related validity is a useful tool for evaluating the effectiveness of a test in a particular context, for example in assessing the ability of a particular test to predict job success. The concept of validity generalisation refers to the extent to which a test may be applied in contexts different from the one for which it was validated. Bearing in mind that it is costly and not always practical to have to validate a test in every context for which it is to be used, particularly if it has already been validated in similar contexts, it has been of concern to organisational psychologists to establish the bounds of test generalisability. Studies conducted in the past have suggested that validity generalisation is not possible (Ghiselli, 1966; Guion, 1965). For example Ghiselli (1966) conducted research in industrial settings which demonstrated that using the same predictors to predict similar criteria for different subjects in comparable settings produced widely varied results. Similarly, great variability was found when validation studies were conducted in educational settings, when the criteria were various school courses (Bennett, Seashore and Wesman, 1974). On the basis of findings such as these, the American Psychological Association regarded tests as being situationally specific, and implemented legislation requiring selection tests to be validated for the

specific contexts in which they were to be used (American Psychological Association, 1975). Until the mid 1970's the situational specificity of psychological requirements was regarded as one of the most serious limitations of organisational psychology (Guion, 1976).

A change in these views on situational specificity came with the work of Hunter, Schmidt and their colleagues. In their initial articles Hunter and Schmidt argue that in most cases psychologists validate their tests on small samples (forty to fifty) of testees (for example, Schmidt, Hunter and Urry, 1976; Schmidt and Hunter 1977; Schmidt, Hunter, Pearlman, and Shane,1979). Hunter and Schmidt suggest that when validity generalisation researchers interpret the results of such studies, they demonstrate faith in an underlying assumption in "the law of small numbers." This is understood to be the the belief that whatever results apply for large sample numbers will also apply for small samples. Hunter and Schmidt reject this notion on the basis that small sample results are highly unstable, resulting in situation specific results. To support their claims Schmidt and Hunter (1978) conducted research on a sample of 10 000 testees drawn from the army. Data was reported on ten predictors that were used to predict success in thirty five jobs, and results indicated highly similar validity coefficients across jobs. Subsequent research was carried out in a similar way on computer programmers (Schmidt, Gast-Rosenberg, and Hunter, 1980), oil industry employees (Schmidt, Hunter and Caplan, 1981), and clerical employees (Pearlman, Schmidt and Hunter, 1980). The researchers found that the other studies yielded similar findings to the first, and concluded that the sources of error disappear in studies with large sample sizes.

Schmidt and Hunter (1981) point out that the most commonly used employment tests have been aptitude or ability tests, such as tests of verbal and quantitative ability, perceptual speed, inductive reasoning, deductive reasoning etc., and have sought to use such tests in their own research. The implication of their findings is that professionally developed tests designed to sample these types of cognitive skills are broadly predictive of performance in both academic and occupational activities (Anastasi, 1982). A further implication of their work is that tests can be validated on large samples of testees, and test users could simply adopt these validities for their specific contexts provided that the job demands are generally the same. An obvious pitfall of such a viewpoint concerns moving from the stance that tests may not be generalised to a belief that a test validated in one setting is valid for all settings .Currently, organisational psychologists are tending to reject the notion of validity generalisation in its extreme form, but tending to accept within - job validity generalisation for a given population group (Muchinsky, 1986; Hartigan and Wigdor, 1989). For example, if a test is validated as a predictor of secretarial job success in one company, it may be taken as a valid predictor of secretarial job success in other companies as well. However, this does not mean that the same test would be a valid predictor of success in non-secretarial office jobs.

3.3.7 **Meta-analysis**

As part of their ongoing research programme, Schmidt, Hunter and their colleagues have sought to show that sources of variance across validity studies need not necessarily only be controlled for by using large samples. They also propose that sources of variance can be controlled for statistically, enabling research findings from studies with small sample numbers to be cumulated. Their work has involved the development and use of a sophisticated statistical technique called meta-analysis.

The term meta-analysis comes from the work of Glass (1976, 1977) and refers to the integrating of empirical data across studies. Although the term has arisen relatively recently, formal methods for combining observations go back much further. Hartigan and Wigdor (1989) point out that integrative statistical techniques were employed in the 1800's to combine different astronomical findings, and in the 1930's to combine significance tests and combine estimates of effects in agricultural experiments. The same authors cite other examples of instances where integrative statistical techniques have been employed over a number of decades. This includes the field of particle physics, combining data from different experiments worldwide; the field of medicine, systematising results of clinical trials; and the behavioural science field, in connection with experiments assessing the effectiveness of psychotherapy, the effects of class size on achievements, and the social psychology of gender differences.

One of the definitive works in the area of meta-analysis is that of Hunter, Schmidt and Jackson (1982) which describes both quantitative and qualitative procedures for integrating findings across studies. These procedures apply not only to simple correlation coefficients, but also to statistical procedures such as regression, canonical correlation and multivariate analysis of variance.

In order to assess the generalisability of test validities across situations, populations of applicants and jobs, the predominant approach to validity generalisation is to separate the "true" validity of a test for given jobs from a number of sources of "artifactual" variance (Hunter et al, 1982). These sources of variance are:

1. Sampling error (i.e. the fact that a given sample is unlikely to be completely representative of the wider population)
2. Differences between studies in range restriction. As has been noted earlier in connection with reliability and validity coefficients, correlations are affected by the relative homogeneity of a given test group (Anastasi, 1982)
3. Unreliability of tests used
4. Unreliability of criteria and methods of evaluating these across studies.

Hunter, Schmidt and their colleagues have developed computational procedures for calculating these primary sources of significance variance (for example, Schmidt and Hunter, 1977; Schmidt, Hunter and Pearlman, 1982; Hunter, Schmidt and Jackson, 1982). These researchers argue that if most of the variability in validities observed across studies can be

accounted for by the artifacts of sampling error, restriction in range and unreliability of test and criterion, it is reasonable to assume that much of the rest can be accounted for by other artifacts such as computational and typographical error. Hunter and Schmidt (1977) propose that if the four primary sources of significance variance can account for 75% of variation, then the remaining 25% may be attributed to other artifacts. The claims made by Hunter,Schmidt and their colleagues concerning meta-analysis have been evaluated in a recent study on the General Aptitude Test Battery ,conducted under the auspices of the American National Research Council.Having critically evaluated each of the steps of meta-analysis in the course of the study,the research committee concluded that the premises of this procedure are statistically sound (Hartigan and Wigdor, 1989).

3.3.8 **Testing and Productivity in Monetary Terms**

Of clear relevance to the use of selection or screening tests in industry is the effect of such tests on general organisational productivity. Some of the original work in this area was carried out by Brogden (1946), who following the work of Taylor and Russell (1939) demonstrated that the expected increase in output is directly proportional to the validity of a test. Other initial work in this area has been carried out by Cronbach and Gleser (1965), but both of these initial studies employed cumbersome procedures.Consequently,Schmidt, Hunter, McKenzie and Muldrow (1979) suggested simplified procedures that are intended to make productivity analysis more viable. Unlike previous research such as Taylor and Russell (1939) where the practical value of a selection procedure has been estimated in terms of the increase of "successful" workers, productivity analysis seeks to estimate the monetary gains that would be realised in using a valid selection instrument.

Using these simplified procedures, Schmidt, Hunter, Pearlman and Shane(1979) have set about estimating the dollar value increase in productivity from using a computer aptitude test to select new computer programmers for the federal government. Their results indicate that when their test (validity 0.79) was compared with random selection, the dollar gain ranged from 97.2 million dollars for a selection ratio of 0.05 to 16.5

million dollars for a selection ratio of 0.80. The gains were estimated over the average expected employment time of 10 years. In other similar studies, Schmidt, Hunter and their colleagues estimated the value of productivity gains resulting from the use of a selection test as: over 32 million dollars per year in the hiring of 2 000 budget analysts (Schmidt and Hunter,1980); 18 million dollars per year for employees of the Philadelphia police department (Schmidt and Hunter, 1981); 16 billion dollars a year for employees of the federal government (Schmidt and Hunter, 1981). Based on research such as this, these authors estimate that optimal test use would have resulted in a benefit of 79.36 billion dollars to employers using the official Employment Service system in 1980 (U.S Department of Labour, 1983, cited Hartigan and Wigdor, 1989), and estimated productivity gains of between 13 and 153 billion dollars in the American economy as a whole due to using ability tests for selection (Hunter and Schmidt, 1982, cited Hartigan and Wigdor, 1989).

As a response to these astonishingly large monetary gains to be expected from test use, a committee under the auspices of the American National Research Council set out to investigate the claims of Hunter and Schmidt (Hartigan and Wigdor, 1989). Tracing the premises of Hunter and Schmidt's productivity calculations, the committee identified a number of fundamental weaknesses, which led them to reject the researchers' estimates of specific dollar gains from test-based selection. On the basis of their work, the committee warn that current scientific knowledge is not sufficient to allow realistic productivity analysis as a result of election practices to be computed (Hartigan and Wigdor, 1989).

CHAPTER 4

COMPARABILITY, TEST BIAS AND THE IMAT

South Africa is a country inhabited by a variety of groups of people, each group sharing in common certain future such as language, history and race (Rhoodie, 1985, cited Taylor and Radford, 1986). Of importance to the following discussion is the fact that where groups share common language, history and socialisation, they are likely to have unique ways of attributing meaning to facts or phenomena (Taylor, 1990). If membership of a particular social group can affect the way the people perceive the world, the question arises as whether membership of different groups can affect scores on a given test. The issue of whether test scores have the same meaning across different groups has generated much research interest. Comparability research focuses on the extent to which the test performance of different groups of people can be compared with the performances of other groups (Poortinga, 1971).

Historically, different population groups in South Africa have been differentiated on the basis of race, with different groups being classified as White, Black, "Coloured" and Asian. This arbitrary racial classifications formed the basis for the discredited ideology of apartheid which has pervaded the social, political and economic structure of South Africa. Despite the fact that apartheid is in the process of being dismantled, the effects of this system will continue to differentiate the opportunities for cognitive development across groups for some time yet, which in the turn is likely to be reflected in psychometric test scores (Taylor and Radford, 1986). Bearing in mind this probable link between apartheid created groups and test performance, the apartheid- classification system for differentiation groups will be used in the study for purposes of criteria discussions.

In the South African literature it is relatively common to find the term "culture" being used somewhat loosely to define a particular population group (e.g. Irvine, 1969; Spence, 1982; Mauer and Retief, 1987; Retief, 1988; Poortinga, 1971). Usually the term "culture" is used synonomously with traditional apartheid categorisation, for example reference may be made to "cross-cultural" research on Blacks, Whites, Coloureds and Asians (Biesheuvel, 1987), or to the "Black culture" or "White culture" (Retief, 1988). A problem with this terminology is that it does not give recognition to the fact that there is no single "White", "Black" or "Coloured" culture. For example there are people who are "White" in South Africa who have very different languages and histories, and undergo different socialisation experiences (eg. English, Afrikaans and Portuguese speaking people). Similarly, there are people who are "Black" who have different languages and histories, and undergo different socialisation experience (eg.Zulu, Indian and Xhosa speaking people). It is possible that sub groups may have unique ways of attributing meaning to events on tests. Therefore for purposes of this discussion and to avoid confusion, the term "culture" will be avoided, and instead the term "ethnic group" will be used. "Ethnic group" will be taken to refer to groups sharing the feature of language, history, and socialisation in common (after Rhoodie,1985; cited Taylor and Radford, 1986).

Reference to the initial history of comparability research in Africa reveals assessment related studies that, more than anything, reflect the beliefs and the prejudices of the researchers. From the early 1900s, studies tended to involve the uncritical application of assessment instruments developed and standardised in Western countries (Retief,1988). The results of these studies were often used to reinforce existing beliefs in the superiority of different groups. A well known example of a comparability study involving the uncritical use of tests is that of the researcher Fick, who in the 1930's assessed the abilities of Black subjects using derivatives of the Binet test for individual assessment, and the American Army Beta test for group testing (Verster, 1987).He used the results to "prove" the superiority of Whites. Increasing criticism (for example Biesheuvel,1943) of uncritical comparability research led to attempts to develop "culture-free" testing in the 1940's and 1950's. In the attempt to develop assessment instruments that were "culture-free", researchers attempted to identify and incorporate elements that were equally familiar to all ethnic groups. However, the results of such studies indicated that even "culture-free" tests were strongly affected by contaminating variables such as literacy and exposure to different types of developmental stimuli (Retief, 1988).

Much work in bringing attention to flaws in comparability research was undertaken by Biesheuvel (e.g. 1943,1949,1952,1958). He stressed the need to give closer attention to the variables of environmental context, such as home life, education and economic status and argued that even careful attempts to adapt the item content of Western tests to a particular culture did not necessarily imply that test scores could be considered comparable. Biesheuvel (1949) argues that in order to test abilities in a given ethnic groups, a sound understanding of the skills and the knowledge of the target population is necessary for the identification of ethnically appropriate items. A related stance is taken by Irvine (1969) who, after reviewing comparability studies involving more than 5 000 people, concluded that it is almost impossible to assess constructs from one ethnic group by using assessment techniques developed in another unless one is certain that the testees involved have experienced the learning conditions necessary for the meaningful interpretation of test scores.

Relatively recent research conducted by Taylor and Radford (1986) presents data indicating that different ethnic groups obtain significantly different mean ability test scores on instruments that are currently being used in the industrial sector in South Africa. There have been various studies of relevance to the phenomenon of different mean scores across different ethnic groups. Verster (1984) compared the performance of White and Black subjects on a battery of twelve cognitive tests and suggested that the Black subjects showed a greater tendency than Whites to trade off speed in favour of accuracy, especially on conceptual tasks, which may provide one clue in explaining test score differences. A study by Van den Berg (1985) on the test score differences between Whites, Asians and "Coloured" school pupils indicated that the test scores were similar across groups where socio-economic status was taken into consideration. This points to the importance of the environmental factors on cognitive development.

In the South African context it is undoubtable that the apartheid structures have unequally affected opportunities for cognitive development across the apartheid created racial groups, which provides another explanation for test score differences. A study by Owen (1986), conducted on a large sample of first year Technicon students revealed significant differences in the patterns of responses of different racial groups on the Senior Aptitude Test Battery. This lead Owen to tentatively conclude that the tests systematically underestimate the abilities of "Coloured", Asian and Black students, due possibly to item bias and the different work tempo that characterised the various groups.

Review of research such as this led J.M.Taylor (1987) to conclude that since test scores do not have the same criterion - referenced meaning for different ethnical groups, there is a strong possibility that unfair labour practices will occur if tests are applied uncritically as selection instruments. It is beyond the scope of this study to review exhaustively cross-ethnic assessment research. Interested readers are referred to Retief (1988). Mauer and Retief (1987), Wober (1975), and Hoorweg (1976) for more comprehensive accounts.

One of the central reasons for the need for comparability research in South Africa is due to the widespread use of psychometric instruments for selection and placement of personnel, which is going to affect increasingly the supply and utilisation of labour (Holburn, 1990). While there is an onus on test users to evaluate their testing practices to ensure that these are fair (Taylor and Radford, 1986; Taylor,1990; Holburn, 1990), it is the responsibility of test constructors to evaluate their instruments for group bias (T.R.Taylor,1987). The remainder of this chapter will be concerned with highlighting some comparability research which has been done on the IMAT. However, firstly it will be useful to consider the concepts of equivalence, bias and fairness.

4.1 EQUIVALENCE AND COMPARABILITY

In order for any two phenomena to be meaningfully compared, it is clearly necessary that they share some feature in common. Failure to have a common feature on which to base the comparison would render the comparison meaningless (Retief, 1988). This raises one of the central issues concerning test score comparability, namely whether the same construct is being assessed across ethnic groups. A second, related issue, concerns whether intergroup differences in test scores reflect real differences in the construct measures, or whether they reflect contaminating variables systematically favouring one group over another.

To serve as a framework for researching these issues, Van der Vijver and Poortinga (1982) propose that phenomena for comparison may be classified into one of four categories of universals, namely: conceptual, functionally equivalent, metrically equivalent and scalar equivalents. These correspond respectively with the four well known forms of measurements: nominal, ordinal, interval and ratio. Phenomena fitting into the "conceptual" category cannot be subject to empirical comparison because they are too abstract (Van der Vijver and Poortinga, 1982). However, it is possible to compare empirically phenomena in any of the remaining levels of category. With regard to tests, Poortinga (1971, 1975) proposes three levels of comparability, corresponding to constructs fitting the functionally equivalent, metrically equivalent and scalar equivalent universals. These are functional equivalence (i.e the requirement that tests qualitatively measure the same attribute in different groups), score equivalence (i.e the requirement that tests measure the attribute on a common scale in the different groups), and item equivalence (i.e that each item in the measuring instrument must be score equivalent - hence each item is treated as a scale in its own right).

While the theory of equivalence is a useful concept, it has limited practical usefulness. T.R.Taylor (1987) suggests that an "infinite regression " problem exists in applying an equivalence model, namely that a score equivalent reference measure is required in order to determine whether other tests or items are score equivalent - but where is one to obtain the reference measure to establish the score equivalence of the reference measure? A more practical way of approaching the problems of test comparability appears to lie in the concept of test bias.

4.2 BIAS

In the context of testing, bias may be understood to refer to influences caused by factors not related to the construct being assessed,resulting in the systematic occurrence of group differences (Retief, 1988; T.R. Taylor, 1987). Test bias may arise from a variety of sources,from individual items to the test as a whole, and from the test administrator to the testing environment (see for example Cole and Means, 1981).Bias can never be eliminated,but steps can be taken to minimise its effects (Taylor and Radford, 1986).

It is possible to distinguish between two main types of bias, namely predictive bias and item bias (T.R.Taylor, 1987). Predictive bias refers to the systematic occurrence of group differences when a test is being used as a predictor against a given criterion. Conversely, item bias refers to the extent to which an item functions differently for different groups. It may be noted that an item which is more difficult for one group is not necessarily biased,but may reflect the effects of generic differences such as age or education (T.R.Taylor, 1987).

For purposes of this discussion it is also important to differentiate the concept of test bias from the concept of "fairness". While test bias refers to influences resulting in the systematic occurrence of group test score differences, fairness refers to the extent to which testing practices favour some groups over others (Holburn, 1990). Hunter and Schmidt (1976) note that fairness is in fact an ethical issue,and that it ultimately rests on competing philosophical perspectives. Consequently, Donald, Veldsman, Donald, Cook, Chemel and Taylor (1990) propose that test users need to develop a comprehensive policy with regard to testing that is consistent with a particular philosophical stance.

It has been mentioned above that it is the responsibility of test constructors to evaluate the instruments that they develop for group biases, and it is the responsibility of test users to evaluate their testing practices to ensure that these are fair. In carrying out these respective responsibilities, test users would be primarily concerned with issues of predictive bias, while test constructors would be primarily concerned with issues of item bias (Taylor, 1990).

Bearing in mind that predictive bias is defined in items of errors of prediction of performance on criteria, it is clear that this concept is closely associated with issues of fairness. It may be noted that a test which is productively biased against a certain group is not inherently unfair, but it is the use to which such a test is put that can be spoken of as fair or unfair. T.R. Taylor (1987) notes that a biased test can be used fairly, and an unbiased test can be used unfairly. Issues of predictive bias and models of fairness are beyond the scope of this study, and interested readers are referred to Jensen (1980), Donald et al (1990), Lautenschlager and Mendoza (1986), J.M.Taylor (1987) and Holburn (1990), for further coverage of these areas.

Concerning item bias, T.R.Taylor (1987) outlines and critiques six bias detection techniques, namely: analysis of variance; transformed item difficulties; item characteristic curve; chi square; log-linear and logit; and regression and partial correlation. Holburn (unpublished manuscript) has applied the logit method of item bias detection to the IMAT performance of a number of samples across Black, White, Asian and Coloured ethnic groups, the results of which will now be considered.

4.3 ITEM BIAS RESEARCH AND THE IMAT

As was mentioned in Section 2.3 above, the IMAT is historically linked to the American developed Otis Quick-Scoring Mental Ability Test. It may therefore be said to reflect Western socialisation values. In order to establish the presence of any biased items in the IMAT, Holburn (unpublished manuscript) performed comparisons between the scores of Black and White samples, using the logit method of bias detection (T.R.Taylor, 1987). The biographical and score distribution details of these various samples are provided in Chapter 6, Table 1.3.

Following the Black-White comparisons, ten items were identified as biased in 50 percent or more of the comparisons. Baseline comparisons between samples of the same race (i.e Black-Black and White-White comparisons) revealed that two of these ten items (items 2 and 4) showed up as biased in the White-White comparison, and were dropped from the list of items that were considered biased. Therefore, from the Black-White comparisons, 8 items were identified as biased, namely items 6 (alphabetical code), 15 (alphabetical reasoning), 16 (alphabetic code), 20 (verbal analogy), 23 (verbal reasoning), 25 (alphabetic series), 27 (alphabetic code) and 28 (alphabetic series).

Following the Coloured-White comparisons, three items were identified as biased. However, two of these (items 2 and 4) also showed evidence of bias in the White-White and Coloured-Coloured comparisons, and were hence dropped from the list of items that were considered biased. Therefore, from the Coloured-White comparisons, only item 20 (verbal analogy) was identified as biased.

Following the Asian-White comparisons, no item showed up as biased (the Asian sample of 14 people was not used in the item bias calculations on account of its small size).

Overall, the results of Holburn (unpublished manuscript) indicate that items based on the alphabet (i.e alphabetic codes and alphabetic series) are biased against Blacks, and that item 20 (verbal analogy) is biased against both Blacks and Coloureds.

Bearing in mind that certain items have been identified as biased, the question arises as to what should be done about this. T.R.Taylor (1987) provides a number of options that are open to the test publisher when bias is detected in a test. These include:

1. Restriction of the use of the test to applications where the bias problem does not occur.

2. Withdrawal of the test with or without the replacement of the test with non-biased items.

3. Warning test users that the test is not comparable across groups, and encouraging test users to employ compensatory selection procedures (e.g Holburn, 1990).

Taylor (1990) indicates that one of the simplest ways that a test user can compensate for the effects of test bias is through the use of separate norms for different groups, and the same normalised cut-off for all groups. Taylor (1990) points out that this practice not only largely overcomes the problem of item bias, but that it may even favour disadvantaged individuals (in that more of them would be selected than would be the case if the sole criterion for selection was "the most effective person for the job").

CHAPTER 5

TABLES OF RESEARCH DATA ON THE IMAT

Tables 1 to 3 summarise the findings of 17 published studies and one unpublished study that have been conducted using the IMAT. Further information on these studies is provided in the review in Chapter 6. For ease of reference three kinds of table have been prepared,numbered 1, 2 or 3. Tables 1.1 to 1.10 provide a general description of the study. They provide information on the author, the purpose of the study,the samples used and the test characteristics of these samples. Studies are presented alphabetically on the basis of the authors' names. Each study is also numbered, and it is by means of these numbers that the respective studies are referred to in the second and third types of tables. Tables 2.1 to 2.4 present correlations between the IMAT and different criteria. Tables 3.1 to 3.5 present correlations between the IMAT and various other tests. Where correlations are provided, significance levels of these correlations are denoted by means of asterisks, which are indicated at the base of the relevant pages. It should be noted that not all of the subjects in a given study wrote all of the tests in that study, and therefore some disparity may be observed in the sample numbers reported in Tables 1,2 and 3.

TABLE 1.1 : DESCRIPTION OF SAMPLE AND TEST CHARACTERISTICS

AUTHOR AND REFERENCE NO.	MEAN SCORE	SD	RELIABILITY (KR 21)	DESCRIPTION	AGE	EDUCATION	TESTING LANGUAGE	YEAR OF TESTING
1. Christierson (1977)	12,47	6,40	0,91 (with Tucker)	Study conducted to determine valid predictors of training success in an instructor course. IMAT used as a predictor. Sample N = 68 (65 White, 5 Coloured) Male construction workers taken on in the Civil Engineering Industry Training Board's basic training course in instructional techniques.	Range 21-63 Mean 37,95	Range Std 5-10 Mean Std 8,44	English and Afrikaans	1976
2. Epstein (1983)	16,20	4,17	0,59 0,74 (with Tucker)	Study conducted to investigate the factor structure of tests theoretically related to mechanical aptitude (including the IMAT), and to determine the extent to which biographical and interest variables may be used to predict performance on the major dimensions of mechanical ability. Sample N = 252 (Black) Male applicants applying for training as technicians.	Range 18-29	Std 8-10	English	1978
3. Erwee (1981)	15,34 15,04 15,18	4,40 4,14 4,20	Not stated	Study conducted to compare cognitive functioning, achievement motivation and vocational preferences amongst male and female university students. IMAT used as an indicator of general cognitive ability. Sample 1 N = 49 (Black) Male first year students, University of Fort Hare. Sample 2 N = 61 (Black) Female first year students, University of Fort Hare. Samples 1 and 2 combined	Mean 20	Matric	English	1979

TABLE 1.2 : DESCRIPTION OF SAMPLE AND TEST CHARACTERISTICS

AUTHOR AND REFERENCE NO.	MEAN SCORE	SD	RELIABILITY (KR 21)	DESCRIPTION	AGE	EDUCATION	TESTING LANGUAGE	YEAR OF TESTING
4. Gieseke (1970)	18,87	5,38	Not stated	Study conducted to predict the ability to learn French, necessary to undergo training in various technical fields. IMAT used as a predictor. Sample N = 40 (26 English, 14 Afrikaans) Male naval ratings	Not stated	Not stated	English and Afrikaans	Approx. 1967
5. Hall (1978)	12,78	4,45	Not stated	Study conducted as pilot research to determine whether the regular counselling service at the NIPR could be successfully adapted for use with Black matriculants. IMAT used primarily as a tool to facilitate vocational counselling. Sample N = 58 (Black) From a population of 78 males and 42 females tested at the NIPR.	Range 18-25	Mainly Matric	English	1977-1978
6. Hall (1980)	18,84	4,45	0,80	Study conducted to determine valid predictors of training success amongst nurses enroled for the General Diploma in nursing. IMAT used as a predictor. Sample N = 277 (White) From a population of 310 female and 2 male student nurses enroled for the general diploma in nursing at the Johannesburg General Hospital.	Range 17-24	Junior Certificate to Matric	English and Afrikaans	1977-1978
7. Halstead (1985)	Not stated	Not stated	Not stated	Study conducted to validate three tests of arithmetical ability (High Level Arithmetical Reasoning Test, Standard Level Arithmetical Reasoning Test and High Level Estimation Test), which were developed to make allowance for the effect of the pocket calculator. The IMAT was used as a cross validation instrument. Sample 1 (Sample 5 in study) N = 91 (Black) Male technikon and technical college students. Sample 2 (Sample 8 in study) N = 136 (Black) Technikon students	Not stated Not stated	First, second or third year technikon Not stated	Not stated	

TABLE 1.3 : DESCRIPTION OF SAMPLE AND TEST CHARACTERISTICS

AUTHOR AND REFERENCE NO.	MEAN SCORE	SD	RELIABILITY (KR 21)	DESCRIPTION	AGE	EDUCATION	TESTING LANGUAGE	YEAR OF TESTING
Halstead (continued)	Not stated	Not stated	Not stated	Sample 3 (Sample 9 in study) (From Epstein), 1983) N = 252 (Black) male applicants for training as technicians.	Range 18-29	Std 8-10	Not stated	Not stated
8. Holburn (unpublished manuscript)				Study conducted to evaluate the item bias in several tests including the IMAT.				
	18,29	4,92	0,8	Sample 1 (Sample 1 in study) N = 206 (Asian) predominantly male applicants for apprenticeships and technical positions in the sugar industry in Natal.	Range 16-37 Mean 20,55	Std 8-10	English	1987-1988
	12,75	4,36	0,75	Sample 2 (Sample 1 in study) N = 208 (Black) predominantly male applicants fort apprenticeship and technical positions in the sugar industry in Natal.	Range 18-33 Mean 23,00	Std 8-10	English	1986-1988
	18,25	4,97	0,8	Sample 3 (Sample 1 in study) N = 102 (Coloured) predominantly male applicants for apprenticeship and technical positions in the sugar industry in Natal.	Range 17-31 Mean 20,06	Std 8-10	English	1981-1988
	19,62	4,58	0,77	Sample 4 (Sample 1 in study) N = 99 (White) predominantly male applicants for apprenticeship and technical positions in the sugar industry in Natal.	Range 17-27 Mean 19,77	Std 8-10	English	1981-1988
	12,19	3,72	0,66	Sample 5 (Sample 2 in study) N = 52 (Black) predominantly male applicants for apprenticeship positions who were selected by a development company in Natal.	Range 18-38 Mean 23,00	Std 8-10	English	1985-1988
	17,14	3,48	0,58	Sample 6 (Sample 3 in study) N = 14 (Asian) male applicants for apprenticeship positions with a large motor company in the Cape.	Range 18-23 Mean 19,50	Std 8-10	English	1989
	12,32	3,62	0,63	Sample 7 (Sample 3 in study) N = 128 (Black) male applicants for apprenticeship positions with a large motor company in the Cape.	Range 17-38 Mean 23,20	Std 8-10	English	1989

TABLE 1.4 : DESCRIPTION OF SAMPLE AND TEST CHARACTERISTICS

AUTHOR AND REFERENCE NO.	MEAN SCORE	SD	RELIABILITY (KR 21)	DESCRIPTION	AGE	EDUCATION	TESTING LANGUAGE	YEAR OF TESTING
Holburn (continued)				Sample 8 (Sample 3 in study)				
	16,33	4,56	0,75	N = 199 (Coloured) male applicants for apprenticeship positions with a large motor company in the Cape.	Range 15-29 Mean 20,70	Std 8-10	Afrikaans	1989
				Sample 9 (Sample 3 in study)				
	15,15	6,22	0,86	N = 74 (White) male applicants for apprenticeship positions with a large motor company in the Cape.	Range 16-25 Mean 19,30	Std 8-10	Afrikaans and English	1989
9. Kendall (Unpublished NIPR study)				(Full information on this study is not available)				
				Sample				
	15,69	4,18	0,59	N = 157 (Coloured) Applicants for apprenticeships at General Motors, Eastern Cape.	Not known	Junior Certificate to Matric	English	1980-1982
10. Latti (1978)				Study conducted to evaluate prevailing selection procedures for clerks, tellers and typists applying for employment with a particular building society. The IMAT was one of the tests being evaluated since it was part of the prevailing selection battery.				
				Sample 1				
	19,06	4,72	Not stated	N = 40 males and 244 females (White) Clerk recruits at the building society.	Range 17-61 Mean 25,32			
				Sample 2				
	18,64	4,17	Not stated	N = 3 males and 289 females (White) Teller recruits at the building society.	Range 16-47 Mean 22,38	English and Afrikaans	Mainly Junior Certificate to matric	Approx. 1976
				Sample 3				
	16,10	4,57	Not stated	N = 63 females (White) Typist recruits at the building society.	Range 17-47 Mean 24,87			
11. Lewis (1980)				Study conducted to establish a test battery suitable for the selection of chargehands (workers supervising the production of mechanical parts). IMAT used as a measure of "general trainability", perceived to be a function of general intelligence.				

TABLE 1.5 : DESCRIPTION OF SAMPLE AND TEST CHARACTERISTICS

AUTHOR AND REFERENCE NO.	MEAN SCORE	SD	RELIABILITY (KR 21)	DESCRIPTION	AGE	EDUCATION	TESTING LANGUAGE	YEAR OF TESTING
Lewis (continued)	10,92	4,78	0,72 0,83 (with Tucker)	Sample N = 79 (Predominantly Afrikaans, some English) Chargehands at Union Carriage and Wagon Company (Pty) Ltd.	Range 24-59 Mean 40	Average Std 8	English and Afrikaans	1979
12. Poortinga (1971)			(Odd-even split-half reliability)	Study conducted to establish the functional equivalence of a series of perceptual scales across the Black and the White culture. The IMAT was used as one of the instruments to help gauge functional equivalence.				
	11,5	6,00	0,78	Sample 1 N = 20 (Black) Male undergraduate students (university not specified)	Range 18-24	Undergraduate - at least one course in Psychology	English	1970
	13,2	5,21	0,62	Sample 2 N = 20 (Black) Female undergraduate students				
	12,3	5,61	0,72	Samples 1 and 2				
	23,1	3,84	0,52	Sample 3 N = 20 (White) Male undergraduate students				
	23,2	3,32	0,72	Sample 4 N = 20 (White) Female undergraduate students	As above	As above	As above	As above
	23,2	3,60	0,44	Samples 3 and 4				
13. Spence (1982)				Study conducted to determine a valid procedure in the selection of Black guidance teachers for primary schools. An attempt was made to see if the IMAT could be used usefully at the educational level equivalent to that of a Black primary school guidance teacher.				
	10,42	3,96	Not stated	Sample 1 N = 237 (Black) 49 male, 188 female primary school teachers in the service of the Johannesburg region of the Department of Education and Training.	Range 20-40	Std 8-10	English	Approx. 1980

TABLE 1.6 : DESCRIPTION OF SAMPLE AND TEST CHARACTERISTICS

AUTHOR AND REFERENCE NO.	MEAN SCORE	SD	RELIABILITY (KR 21)	DESCRIPTION	AGE	EDUCATION	TESTING LANGUAGE	YEAR OF TESTING
Spence (continued)				Sample 2				
	16,57	4,81	0,73 0,814 (with Tucker)	(From Visser, 1978) N = 507 (Black) male and female first year students, University of Fort Hare.	Not stated	Matric	English	1977-1978
14. Taylor (1983)				Study conducted to determine a valid procedure in the identification of potential engineering technicians from the Black population group. The IMAT was one of the tests assessed for its effectiveness in predicting successful course results.				1980-1981
			0,68	Total Sample				
	18,5	4,6	0,81 (with Tucker)	N = 149 (Black) All first year students enroled at the Mangosuthu Technikon (Umlazi) in the Engineering Department.	Mean 22	Matric	English	
				Group 1 (N = 118)				
	19,1	4,1	Not stated	Those students who completed first semester (T1) course examinations.	As above	As above	As above	As above
				Group 2 (N = 31)				
	16,0	5,5	Not stated	Those students who dropped out before T1 examinations.				
15. Taylor, Werbeloff and Ebertsohn (1982)				Study conducted to validate the mainframe-computerised, English and Afrikaans versions of the IMAT against their pencil-and-paper counterparts.				
				Sample 1A (Computerised)				
	17,94	5,41	0,78	N = 150 (White) Predominantly male (between 61 and 66 percent) ESCOM job applicants for clerical and technical positions.	Mean 25	Std 8-10	English	Approx. 1980
				Sample 1B (Paper-and-Pencil)				
	18,18	5,46	0,79	N = 190 (White) Predominantly male (between 61 and 66 percent) ESCOM job applicants for clerical and technical positions.	Mean 25	Std 8-10	English	Approx. 1980
				Sample 2A (Computerised)				
	15,57	5,43	0,77	N = 150 (White) Predominantly male (between 61 and 66 percent) ESCOM job applicants for clerical and technical positions.	Mean 25	Std 8-10	Afrikaans	Approx. 1980

TABLE 1.7 : DESCRIPTION OF SAMPLE AND TEST CHARACTERISTICS

AUTHOR AND REFERENCE NO.	MEAN SCORE	SD	RELIABILITY (KR 21)	DESCRIPTION	AGE	EDUCATION	TESTING LANGUAGE	YEAR OF TESTING
Taylor, Werbeloff and Ebertsohn (continued)				<u>Sample 2B (Paper-and-Pencil)</u>				
	15,68	5,65	0,79	N = 213 (White) Predominantly male (between 61 and 66 percent) ESCOM job applicants for clerical and technical positions.	Mean 25	Std 8-10	Afrikaans	Approx. 1980
				<u>Sample 3A (Computerised)</u>				
	6,22 (Verbal items)	Not stated	Not stated	N = 150 (White) Predominantly male (between 61 and 66 percent) ESCOM job applicants for clerical and technical positions. Mean scores on the 12 "verbal" items of the IMAT (analogies and logical problems) contrasted with the mean scores on the 18 "non-verbal" items (series and codes).	Mean 25	Std 8-10	English	Approx. 1980
	11,72 (non-verbal items)	-	-					
				<u>Sample 3B (Computerised)</u>				
	5,09 (verbal items)	-	-	N = 150 (White) Predominantly male (between 61 and 66 percent) ESCOM job applicants for clerical and technical positions. Mean scores on the 12 "verbal" items of the IMAT (analogies and logical problems) contrasted with the mean scores on the 18 "non-verbal" items (series and codes).	Mean 25	Std 8-10	Afrikaans	Approx 1980
	10,48 (non-verbal items)	-	-					
				<u>Sample 4A (Computerised)</u>				
	6,75 (verbal items)	-	-	N = 75 (White) Predominantly male ESCOM job applicants for technical positions. Mean scores on the 12 "verbal" items of the IMAT (analogies and logical problems) contrasted with the mean scores on the 18 "non-verbal" items (series and codes).	Mean 25	Std 8-10	English	Approx. 1980
	12,63 (non-verbal items)	-	-					
	19,37 (total)	-	-					

TABLE 1.8 : DESCRIPTION OF SAMPLE AND TEST CHARACTERISTICS

AUTHOR AND REFERENCE NO.	MEAN SCORE	SD	RELIABILITY (KR 21)	DESCRIPTION	AGE	EDUCATION	TESTING LANGUAGE	YEAR OF TESTING
Taylor, Werbeloff and Ebertsohn (continued)				Sample 4B (Computerised)				
	5,41 (verbal items)	-	-	N = 75 (White) Predominantly male ESCOM job applicants for technical positions. Mean scores on the 12 "verbal" items of the IMAT (analogies and logical problems) contrasted with the mean scores on the 18 "non-verbal" items (series and codes).	Mean 25	Std 8-10	Afrikaans	Approx. 1980
	10,56 (non-verbal items)	-	-					
	15,97 (total)	-	-					
				Sample 5A (Computerised)				
	5,55 (verbal items)	-	-	N = 75 (White) Predominantly male ESCOM job applicants for clerical positions. Mean scores on the 12 "verbal" items of the IMAT (analogies and logical problems) contrasted with the mean scores on the 18 "non-verbal" items (series and codes).	Mean 25	Std 8-10	English	Approx. 1980
	10,67 (non-verbal items)	-	-					
	16,21 (total)	-	-					
				Sample 5B (Computerised)				
	4,65 (verbal items)	-	-	N = 75 (White) Predominantly male ESCOM job applicants for clerical positions. Mean scores on the 12 "verbal" items of the IMAT (analogies and logical problems) contrasted with the mean scores on the 18 "non-verbal" items (series and codes).	Mean 25	Std 8-10	Afrikaans	Approx. 1980
	10,20 (non verbal items)	-	-					
	14,85 (total)	-	-					

TABLE 1.9 : DESCRIPTION OF SAMPLE AND TEST CHARACTERISTICS

AUTHOR AND REFERENCE NO.	MEAN SCORE	SD	RELIABILITY (KR 21)	DESCRIPTION	AGE	EDUCATION	TESTING LANGUAGE	YEAR OF TESTING
Taylor, Werbeloff and Ebertsohn (continued)	Not stated	Not stated	Not stated	Sample 6A (Paper-and-Pencil) N = 91 (White) Predominantly male ESCOM job applicants for technical positions. IMAT results correlated with other tests.	Mean 25	Std 8 - 10	English	Approx. 1980
	Not stated	Not stated	Not stated	Sample 6B (Paper-and-Pencil) N = 111 (White) Predominantly male ESCOM job applicants for technical positions. IMAT results correlated with other tests.	Mean 25	Std 8-10	Afrikaans	Approx. 1980
	Not stated	Not stated	Not stated	Sample 7A (Paper-and-Pencil) N = 99 (White) Predominantly male ESCOM job applicants for clerical positions. IMAT results correlated with other tests.	Mean 25	Std 8-10	English	Approx. 1980
	-	-	-	Sample 7B (Paper-and-Pencil) N = 102 (White) Predominantly male ESCOM job applicants for clerical positions. IMAT results correlated with other tests.	Mean 25	Std 8-10	Afrikaans	Approx. 1980
16. Visser (1978)				Study conducted to prepare for a selection programme and careers counselling-programme for Black students, and to validate NIPR tests (including the IMAT) against a criterion of success at university.				
	16,12	4,74	0,692 0,806 (with Tucker)	Sample 1 N = 132 (from a sample of 163 males and 86 females) Black first year students, University of Fort Hare.	Not stated	Matric	English	1977
	16,73	4,82	Not stated	Sample 2 N = 363 (from a sample of 238 males and 135 females) Black first year students, University of Fort Hare.	Not stated	Matric	English	1978
	16,57	4,81	0,703 0,814 (with Tucker)	Samples 1 and 2 combined N = 507	Not stated	Matric	English	1977 + 1978

TABLE 1.10 : DESCRIPTION OF SAMPLE AND TEST CHARACTERISTICS

AUTHOR AND REFERENCE NO.	MEAN SCORE	SD	RELIABILITY (KR 21)	DESCRIPTION	AGE	EDUCATION	TESTING LANGUAGE	YEAR OF TESTING
17. Visser (1980)				Study conducted to investigate the implications of electroencephalographic (EEG) assessments in the vocational guidance of NIPR clients with suspected brain abnormalities. The IMAT was one of several tests used to compare the cognitive performance of people referred for EEG examination with that of a "normal" population.				
	15,01 (20)	4,54	Not stated	<u>TOTAL SAMPLE</u> N = 68 (58 English, 10 Afrikaans) 61 males, 7 females. Subjects were the total number of people who went to the NIPR for counselling, and who were referred for EEG on the counsellors' initiatives.	Mean 19,98	School and University	English and Afrikaans	1965-1978
	16,67	4,23	Not stated	<u>Clinical Group</u> N = 26. Cases with known and treated epilepsy, severe head injury and congenial or long standing brain damage.				
	15,33	5,13	Not stated	<u>Borderline Group</u> N = 19. Cases reporting fainting, isolated blackouts or epileptogenic episodes.				
	13,63	4,44	Not stated	<u>Behaviour Problem Group</u> N = 23. Cases with school and learning problems and behaviour difficulties, sometimes of a psychopathic nature.				
	16,43	4,88	Not stated	<u>Comparison Group</u> N = 205. South African Railways male clerical staff tested for purposes of standardizing the Intermediate Battery.	Mean 19,25	Std 7-9	English	1964-1965
18. Werbeloff and Taylor (1982)				Study conducted to validate the High Level Figure Classification Test (HLFCT) as a selection instrument. The IMAT was used as one of several tests to assess the concurrent validity of the HLFCT.				
				<u>Sample 1</u> (Sample vii in study) N = 263 (Black) Male applicants for training as technicians for electronic and electromechanical equipment. IMAT results correlated with end of year performance after course.	Range 17-29	Range 8-10 yrs	English	1980

TABLE 1.11 : DESCRIPTION OF SAMPLE AND TEST CHARACTERISTICS

	MEAN SCORE	SD	RELIABILITY (KR 21)	DESCRIPTION	AGE	EDUCATION	TESTING LANGUAGE	YEAR OF TESTING
Werbeloff and Taylor (continued)	Not stated	Not stated	Not stated	Sample 2 (Sample viii in study) N = 199 (Race and sex not stated) First year students of Mabopani East Technikon IMAT results correlated with other tests.	Not stated	Not stated	Not stated	1982

TABLE 2.1 : CORRELATION OF IMAT WITH CRITERIA (N in Brackets)

Reference	Civil Engineering Industry Training Board Appraised	NIPR Checklist	NIPR Global rating	Course knowledge test	French vocabulary and grammar	French comprehension	French fluency	French accent	Resignation before end of nursing course	Successful achievement of nursing qualification	Examination Results					
											Digital systems	Electrical engineering	Industrial technology	Engineering mathematics	Communication	Average of above
1. Sample	*** 0,47 (48)	** 0,41 (44)	*** 0,55 (44)	*** 0,74 (37)												
4. Sample					*** 0,449 (40)	*** 0,592 (40)	*** 0,441 (40)	0,230 (40)								
6. Sample									Not significant	Not significant						
7. Sample 1											0,27 (20-32)	-0,04 (20-32)	** 0,51 (20-32)	0,12 (20-32)	* 0,46 (20-32)	0,15 (20-32)

* Significant at the 5 percent level
** Significant at the 1 percent level
*** Significant at the 0,5 percent level

TABLE 2.2 : CORRELATION OF IMAT WITH CRITERIA (N in Brackets)

Reference	Examination Results						Performance rating (3 months)	Performance rating (12 months)	Global rating	Verbal communication	Activity planning	Written communications	Information recording	Financial awareness	Supervisory ability	Safety	Co-operativeness	Leadership	Initiative	Learning ability	Originality	Theoretical knowledge	General efficiency
	Preliminary Mathematics	Chemistry	Applied Science	Preliminary drawing	Communication	Average of above																	
7. Sample 2	*** 0,38 (136)	*** 0,37 (136)	*** 0,26 (136)	0,16 (136)	*** 0,49 (136)	*** 0,42 (136)																	
10. Sample 1							*** 0,32 (284)	-0,04 (284)															
Sample 2							* 0,19 (292)	0,02 (292)															
Sample 3							* 0,37 (63)	-0,12 (63)															
11. Sample									0,003 (64)	* 0,28 (64)	* 0,25 (64)	* 0,33 (64)	0,066 (64)	-0,58 (64)	0,011 (64)	-0,077 (64)	0,067 (64)	-0,044 (64)	-0,097 (64)	0,095 (64)	-0,025 (64)	0,029 (64)	-0,12 (64)

* Significant at the 5 percent level
** Significant at the 1 percent level
*** Significant at the 0,5 percent level

TABLE 2.3 : CORRELATION OF IMAT WITH PREDICTOR CRITERIA (N in Brackets)

Reference	Practical ability	Drawing interpretation	Fault finding	School Results					Pre-Tech Results						First Term Results				Half year examination marks								End of year exam marks			
				Matric English	Matric Afrikaans	Home language	Mathematics	Science	Mathematics	Chemistry	Applied Science	Technical drawing	Communications	Pre Tech Average	No of subjects passed	T1 drawing	T1 mathematics	T1 communications	Mathematics	Private law	Statistical methods	Physics	English	Chemistry	Economics	Psychology	Chemistry	Mathematics	Physics	Biology
11. (continued)	-0,073 (64)	0,18 (64)	-0,12 (64)																											
14. Sample				* 0,19 (132)	* 0,22 (124)	0,14 (141)	** 0,23 (124)	0,06 (120)	*** 0,42 (144)	*** 0,38 (144)	** 0,29 (146)	0,18 (146)	*** 0,48 (146)	*** 0,45 (146)	0,10 (118)	0,01 (125)	* 0,20 (118)	*** 0,52 (101)												
16. Sample 1																			* 0,51 (19)	-0,34 (12)	0,30 (16)	0,20 (16)	0,03 (8)	* 0,44 (17)	0,21 (17)	0,15 (15)	** 0,39 (50)	0,30 (41)	0,37 (41)	-0,11 (22)

* Significant at the 5 percent level
** Significant at the 1 percent level
*** Significant at the 0,5 percent level

TABLE 2.4 : CORRELATION OF IMAT WITH CRITERIA (N in Brackets)

Reference	Electronics	Applied mechanics	Typewriting	Adding machine test	Self study
18. Sample 1	*** 0,49 (66)	* 0,32 (66)	* 0,24 (66)	* 0,29 (66)	*** 0,50 (66)

* Significant at the 5 percent level
** Significant at the 1 percent level
*** Significant at the 0,5 percent level

TABLE 3.1(a) : CORRELATION OF IMAT WITH OTHER TESTS

REFERENCE	TESTS (ALPHABETICAL)			
	Arithmetical Problems (Intermediate)	Arithmetical Reasoning (Standard)	Blox	Computation (Intermediate)
1. Sample	*** 0,75 (48)			
2. Sample	*** 0,614 (136)			
14. Sample	*** 0,62 (149)			
15. Sample 5A	*** 0,68 (37)			
16. Sample 2	*** 0,48 (163)			
2. Sample		*** 0,49 (75)		
7. Sample 1		*** 0,421 (64-91)		
Sample 2		*** 0,565 (136)		
14. Sample		*** 0,56 (136)		
18. Sample 1		*** 0,37 (263)		
Sample 2		*** 0,46 (199)		
2. Sample			0,02 (75)	
7. Sample 1			0,068 (20-32)	
Sample 2			*** 0,262 (136)	
11. Sample			*** 0,569 (62)	

* Significant at the 5 percent level
** Significant at the 1 percent level
*** Significant at the 0,5 percent level

TABLE 3.1(b) : CORRELATION OF IMAT WITH OTHER TESTS

REFERENCE	TESTS (ALPHABETICAL)			
	Arithmetical Problems (Intermediate)	Arithmetical Reasoning (Standard)	Blox	Computation (Intermediate)
12. Samples 1 and 2			*** 0,59 (40)	
Samples 3 and 4			* 0,36 (40)	
14. Sample			*** 0,32 (149)	
15. Sample 4A			*** 0,62 (97)	
Sample 4B			*** 0,54 (96)	
Sample 6A			*** 0,59 (91)	
Sample 6B			*** 0,43 (111)	
16. Sample 1			0,30 (38)	
18. Sample 1			** 0,29 (263)	
Sample 2			** 0,38 (199)	
10. Sample 1				*** 0,50 (284)
Sample 2				*** 0,52 (292)
16. Sample 1				*** 0,53 (144)
Sample 2				*** 0,58 (362)

* Significant at the 5 percent level
** Significant at the 1 percent level
*** Significant at the 0,5 percent level

TABLE 3.2 : CORRELATION OF IMAT WITH OTHER TESTS

REFERENCE	TESTS (ALPHABETICAL)							
	Concept identification	Deductive reasoning test (Intermediate)	Dominoes	Embedded Problems Test	Estimations Test (High level)	F Test	Fault Finding Test	Figure Analogy Test
16. Sample 2	*** 0,45 (204)							
7. Sample 2		*** 0,495 (136)						
16. Sample 2			*** 0,61 (200)					
7. Sample 1				0,304 (20-32)				
7. Sample 1					*** 0,374 (64-91)			
2. Sample						0,02 (75)		
7. Sample 2						*** 0,354 (136)		
15. Sample						*** 0,388 (149)		
18. Sample 1						*** 0,27 (263)		
2. Sample							0,03 (75)	
16. Sample 2								*** 0,56 (201)
2. Sample 2								

* Significant at the 5 percent level
** Significant at the 1 percent level
*** Significant at the 0,5 percent level

TABLE 3.3(a) : CORRELATION OF IMAT WITH OTHER TESTS

REFERENCE	TESTS (ALPHABETICAL)										
	Figure Classification Test (Standard Level)	Figure Classification Test (High Level)	Foreign Language Memory Test	General Science Test	Gottschaldt Figures Test	H Test	Lees Begrip	Mechanical Comprehension Test	Meganiese Insig	O'Connor Finger Dexterity	Poppelreuter
2. Sample	*** 0,50 (75)										
16. Sample 2	*** 0,56 (145)										
18. Sample 1		*** 0,60 (263)									
Sample 2		*** 0,48 (199)									
4. Sample			*** 0,59 (40)								
7. Sample 1				** 0,339 (64-91)							
2. Sample					** 0,31 (64)						
7. Sample 2					*** 0,345 (136)						
11. Sample					*** 0,40 (58)						
14. Sample					*** 0,37 (149)						
2. Sample						0,10 (75)					
7. Sample 2						** 0,357 (136)					
14. Sample						*** 0,41 (149)					
18. Sample 1						*** 0,25 (263)					
15. Sample 5B							*** 0,53 (82)				
Sample 7B							*** 0,55 (102)				

* Significant at the 5 percent level ** Significant at the 1 percent level *** Significant at the 0,5 percent level

TABLE 3.3(b) : CORRELATION OF IMAT WITH OTHER TESTS

REFERENCE	TESTS (ALPHABETICAL)										
	Figure Classification Test (Standard Level)	Figure Classification Test (High Level)	Foreign Language Memory Test	General Science Test	Gottschaldt Figures Test	H Test	Lees Begrip	Mechanical Comprehension Test	Meganiese Insig	O'Connor Finger Dexterity	Poppelreuter
2. Sample								** 0,25 (75)			
7. Sample 1								** 0,526 (64-91)			
Sample 2								** 0,315 (136)			
14. Sample								** 0,35 (149)			
15. Sample 4A								*** 0,54 (97)			
Sample 6A								*** 0,65 (91)			
18. Sample 1								*** 0,41 (263)			
15. Sample 4B									*** 0,56 (96)		
Sample 6B									*** 0,40 (111)		
2. Sample										0,09 (75)	
2. Sample											* 0,21 (75)

* Significant at the 5 percent level
** Significant at the 1 percent level
*** Significant at the 0,5 percent level

TABLE 3.4(a) : CORRELATION OF IMAT WITH OTHER TESTS

REFERENCE	TESTS (ALPHABETICAL)							
	Ravens Matrices	Ravens Matrices (Advanced)	Reading Comprehension (Intermediate)	Rekenkundige Probleme	Spelling (Intermediate)	Soek-die-Fout - Accuracy (Intermediate)	Soek-die-Fout - Speed (Intermediate)	Spot-the-Error - Accuracy (Intermediate)
3. Samples 1 and 2	*** 0,33 (110)							
11. Sample	*** 0,66 (64)							
14. Sample	*** 0,50 (144)							
16. Sample 1	*** 0,54 (134)							
Sample 2	*** 0,61 (352)							
Samples 1 and 2	*** 0,59 (486)							
12. Samples 1 and 2		*** 0,73 (40)						
Samples 3 and 4		0,29 (40)						
3. Samples 1 and 2			*** 0,26 (110)					
10. Sample 1			*** 0,45 (284)					
15. Sample 5A			*** 0,69 (37)					
Sample 7A			*** 0,64 (99)					
16. Sample 1			*** 0,26 (144)					
Sample 2			*** 0,40 (363)					
Samples 1 and 2			*** 0,59 (486)					

* Significant at the 5 percent level ** Significant at the 1 percent level *** Significant at the 0,5 percent level

TABLE 3.4(b) : CORRELATION OF IMAT WITH OTHER TESTS

REFERENCE	TESTS (ALPHABETICAL)							
	Ravens Matrices	Ravens Matrices (Advanced)	Reading Comprehension (Intermediate)	Rekenkundige Probleme	Spelling (Intermediate)	Soek-die-Fout - Accuracy (Intermediate)	Soek-die-Fout - Speed (Intermediate)	Spot-the-Error - Accuracy (Intermediate)
15. Sample 5B				*** 0,64 (82)				
Sample 7B				*** 0,68 (102)				
9. Sample 3					*** 0,31 (63)			
15. Sample 5B						0,19 (82)		
Sample 7B						0,18 (102)		
15. Sample 5B							*** 0,32 (82)	
Sample 7B							*** 0,29 (102)	
9. Sample 1								* 0,17 (284)
Sample 2								0,11 (292)
Sample 3								0,22 (63)
15. Sample 5A								-0,03 (37)
Sample 7A								0,43 (99)

* Significant at the 5 percent level
** Significant at the 1 percent level
*** Significant at the 0,5 percent level

TABLE 3.5(a) : CORRELATION OF IMAT WITH OTHER TESTS

REFERENCE	TESTS (ALPHABETICAL)			
	Spot-the-Error - Speed (Intermediate)	Technical Reading Comprehension	Tegniese Lees Begrip	Vocabulary (Intermediate)
10. Sample 1	* 0,18 (284)			
Sample 2	*** 0,28 (292)			
Sample 3	*** 0,38 (63)			
15. Sample 5A	0,25 (37)			
Sample 7A	*** 0,40 (99)			
1. Sample		*** 0,72 (48)		
2. Sample		*** 0,40 (75)		
4. Sample		*** 0,43 (40)		
7. Sample 2		*** 0,543 (136)		
14. Sample		*** 0,54 (149)		
15. Sample 4A		*** 0,56 (97)		
Sample 6A		*** 0,69 (91)		

* Significant at the 5 percent level ** Significant at the 1 percent level *** Significant at the 0,5 percent level

TABLE 3.5(b) : CORRELATION OF IMAT WITH OTHER TESTS

REFERENCE	TESTS (ALPHABETICAL)			
	Spot-the-Error - Speed (Intermediate)	Technical Reading Comprehension	Tegniese Lees Begrip	Vocabulary (Intermediate)
18. Sample 1		*** 0,42 (263)		
Sample 2		*** 0,28 (199)		
15. Sample 4B			*** 0,61 (96)	
Sample 6B			*** 0,62 (111)	
3. Samples 1 and 2				*** 0,28 (110)
16. Sample 1				* 0,18 (144)
Sample 2				** 0,19 (363)
Samples 1 and 2				*** 0,18 (507)

* Significant at the 5 percent level
** Significant at the 1 percent level
*** Significant at the 0,5 percent level

CHAPTER 6

REVIEW OF RESEARCH STUDIES ON THE IMAT

It has been noted in Sections 3.1.1 and 3.3.7 above that small sample correlation studies are subject to a significant degree of potential error variance. Hunter, Schmidt and Jackson (1982) define "small sample size" for correlational studies as "all studies with less than a thousand persons", and recommend that "confidence intervals give a correct picture of the extent of uncertainty that surrounds results computed from small-sample studies" (p. 24). It may be noted that none of the studies reviewed come close to a subject sample size of a thousand, and they are therefore all subject to the limitations of small-sample interpretation. Consequently, in the review, the confidence intervals of key correlation coefficients will be provided. Should readers wish to obtain confidence intervals of correlations, they are referred to statistical tables such as Neave (1978).

The different studies on the IMAT will be reviewed in the alphabetical order of the author's names (i.e as presented in Tables 1.1 to 1.11 in Chapter 5). Details of each of the samples used in the studies can be found in the tables. In this chapter the individual studies will be considered. General trends and considerations will be provided in Chapter 7.

Christierson (1977) conducted a study to determine valid predictors of training success on the Civil Engineering Industry Training Board's basic training course in instructional techniques. Along with the Intermediate Arithmetical Problems Test (IAPT) and the Technical Reading Comprehension Test (TRCT), the IMAT was hypothesised to validly predict training success on the instructors' course. Determinants of training success were assessed at the end of the training course by the following measures: a course knowledge test (developed by the NIPR containing 26 multiple choice items covering the main points of the course content); a check list (a 37 item instrument designed by the NIPR to give an objective measurement of what the trainee actually does or does not do during a session of instructional performance); a global rating (done prior to the scoring of the check list - a rater's personal judgment of the trainees performance on a scale of five categories ranging from "very poor" to "very good"); and the Civil Engineering Industry Training Board Appraisal - CEITBA (an instrument of 24 items evaluating a trainee's performance as an instructor on a 10-point scale).

It was shown that the IMAT correlated relatively highly with the IAPT (0,75, with a 95 percent confidence interval of $0,56 \leqslant r \leqslant 0,83$). The internal consistency of items on all three tests was high (above 0,9 using the Kuder and Richardson formula with Tucker's correction). Although the IMAT correlated relatively highly with the "training success" criteria (between 0,41 and 0,74 - see Table 2.1), a stepwise multiple regression (which was carried out in addition to the basic Pearson correlations) indicated that neither it, nor the TRCT did much to improve the predictive performance of the IAPT. The correlation between the primary criterion - the CEITBA and the IMAT was 0,47 (with a 95 percent confidence interval of $0,14 \leqslant r \leqslant 0,70$). Conversely the correlation between the CEITBA and the IAPT was 0,61 (with a 95 percent confidence interval of $0,43 \leqslant r \leqslant 0,74$). The addition of the TRCT and the IMAT together only increased the multiple correlation to 0,63. As a result it was recommended that the IAPT alone be used for the

It would appear unusual that a test designed to assess arithmetical problem solving ability should have correlated higher with a criterion of success on a training course in instructional techniques than the IMAT, since the training course in instructional techniques required general reasoning abilities such as the ability to apply basic instructional techniques, to make use of visual aids, to demonstrate the task to be taught, and to evaluate and record trainees' progress (Christierson, 1977). The IMAT will be seen to correlate moderately to highly with a wide variety of tests suggesting that it is a good test of general reasoning ability, which amongst other things should include predicting success on a criterion of learning instructional techniques. It should be noted that the correlations of both the IAPT and IMAT with the criterion are the subject to fairly wide confidence intervals reflecting the small sample sizes.

Christierson (1977) provides relatively comprehensive information on the subjects and their test statistics, and the instruments used in the study which would allow for replication of the study. There is some question over including the "Coloured" subjects with White subjects, given the comments made in Chapter 4. Ideally the study should have checked for significant differences in predictors and criteria for the two groups, to establish whether separate validation studies were necessary.However it is noted that only five "Coloured" subjects appeared to be available.

Epstein (1985) conducted a study to investigate the factor structure of a number of tests theoretically related to mechanical aptitude on a sample of Black trainee technicians.The IMAT was one of the tests used in the study.Factor analysis (using an oblique rotation) of the performances of the sample on the various tests revealed four factor clusters. Interpretation of each factor was primarily based upon tests with high loadings, that is, with loadings above 0,30 (loadings between 0,20 and 0,30 were interpreted more tentatively, and loadings below 0,20 were not interpreted).

The IMAT was found to have a moderate loading on Factor II (0,29) and a fairly strong loading on Factor III (0,46). Factor II is shown to be best defined by the following tests: Technical Reading Comprehension (0,50); Gottschaldt Embedded Figures Test (0,38), and Arithmetical Reasoning (0,36). Epstein suggests that tests which are high on this factor all appear to require "analytic thinking and the processing of information in a series of logical steps" (1985, p.50). In addition to the IMAT, Factor III is shown to be best defined by the Figure Classification Test (0,54), with a lower loading on the Standard Level Arithmetic Reasoning Test (0,28). The author identifies Factor III as a "conceptual reasoning" factor, and states that "Tests which are high on this factor require flexible thinking and the ability to approach a new task with an open mind" (Epstein, 1985, p.53).

Commenting on a lower-than-expected loading of the IMAT on Factor II (considered to be an analytical thinking factor), the author suggests that this may be ascribed at least partly to the high verbal component of the test. This suggestion is based on Werbeloff and Taylor's (1982) assertion that Blacks who are taking the test are doing so in a second language, and as a result may have difficulty understanding what is required of them in certain items, or may have to translate material into their own language before performing logico-verbal manipulations.

Commenting on the higher-than-expected loading of the IMAT on Factor III, Epstein (1985) states that it is likely that this reflects differences in the hypothesis formulating approach of the specific sample to the items of the IMAT.

Epstein notes that in general the reliabilities of the tests in the test battery (including the IMAT) fall short of the minimal 0,75 level of reliability at which test constructors aim (Werbeloff and Taylor, 1982).She attributes this to the relative homogeneity of the test sample, which tends to depress reliability values (Nunnally, 1967).

An intercorrelation matrix between the IMAT and the other tests in the study revealed that the IMAT correlated highest (at the 0,1 percent level) with the High Level Figure Classification Test, the Standard Level Arithmetic Reasoning Test, and the Technical Reading Comprehension Test. Respective correlations and 95 percent confidence intervals for these tests were $r = 0,50$ $(0,37 \leqslant r \leqslant 0,65)$; 0,49 $(0,36 \leqslant r \leqslant 0,64)$; 0,40 $(0,20 \leqslant r \leqslant 0,49)$. This again indicates the overlap in factor domain measured by the IMAT.

In terms of limitations of this study, as with the other studies reviewed, the relatively small sample ($n = 75$) places limitations on the generalisability of the results. Secondly, since the sample was relatively highly selected, restriction in range would probably have reduced the correlations between variables. Thirdly, the author notes that due to the small sample size, many variables which could have affected the test scores(eg.biographical factors),could not be taken into account in the computations.

Erwee (1981) undertook research to compare cognitive functioning, achievement motivation and vocational preferences between male and female Black first year university students at the University of Fort Hare. A distinction was not made between groups in term of ethnicity or degree being studied, as the sample was considered too small to do this. The study was intended to test a number of hypotheses including:

1. Black male students would obtain higher scores than Black females on the IMAT;

2. a positive relationship would be found to exist between various dimensions of achievement motivation (measured by The Achievement Motivation Questionnaire - comprehensively discussed by Pottas et al, 1980), and aspects of cognitive functioning measured by the IMAT, Intermediate Reading Comprehension Test, Intermediate Vocabulary Test, and Raven's Progressive Matrices Test.

No significant difference was found to exist between the male and female samples on the IMAT, leading the author to reject this hypothesis. Low correlations were found to exist between the measured dimensions of achievement motivation and performances on the cognitive tests including the IMAT.

The author tested the IMAT scores for skewness and found these to be negatively skewed, but not at a significant level.

Correlations between the IMAT, and the Reading Comprehension, Vocabulary and Ravens Progressive Matrices results were found to be significant, supporting similar findings by Visser (1978) and Crawford-Nutt (1977). It was postulated by Erwee that a verbal language link exists between the IMAT and the Reading Comprehension and Vocabulary tests, and that the correlation between the IMAT and the Ravens' test may reflect "...the type of common language ie. the figures and diagrams used to reflect the problems in their tests" (Erwee, 1981,p.41).

Again a problem in this study concerns the relatively small sample size,which made it impossible to compare correlation patterns of the various ethnic sub-groups,or different faculties. Internal consistency figures were unfortunately not reported for the samples tested on the IMAT.

In an attempt to predict ability to learn a foreign language amongst a sample of English and Afrikaans Naval recruits, Gieseke (1970) hypothesised that the following would be valuable predictors: verbal fluency in the second official South African language; scholastic achievement in languages; ability to memorise foreign words, and general intellectual ability. Ability to memorise foreign words was assessed by means of an NIPR-developed instrument, the Foreign Language Memory Test (FLMT). General intellectual ability was equated with "trainability", and was assessed by the IMAT. The ability of subjects to comprehend technical and scientific data in their home language (measured by the NIPR-developed Technical Scientific Reading Comprehension Test - TSRCT) was taken as an indication of ability to cope with similar data in a new language. Criterion data consisted of assessments made on a five point scale at the end of an intensive eight-week course in French. The assessments were made on the following factors: grammar and vocabulary; comprehension; fluency, and accent.

It was found that the IMAT correlated significantly at the 0,1 percent level with the grammar and vocabulary assessment ($r=0,45$, with a 95 percent confidence interval of $0,16 \leqslant r \leqslant 0,70$); at the 0,1 percent level with the comprehension assessment ($r=0,60$, with a 95 percent confidence interval of $0,38 \leqslant r \leqslant 0.67$), and at the 1 percent level with the fluency assessment ($r=0,44$, with a 95 percent confidence interval of $0,20 \leqslant r \leqslant 0,69$). The IMAT was found not to correlate significantly with ability to pick up a French accent. This is to be expected since ability to pick up an accent is unlikely to be a function of reasoning ability. Inter-correlations between the three tests used in the study indicated coefficients of 0,59 between the IMAT and FLM (significant at the 0,1 percent level) and 0,43 between the IMAT and the TSRCT (significant at the 1 percent level).

Several criticisms may be raised with regard to this study.Once again the sample size was relatively small. Secondly, the sample comprised both English and Afrikaans subjects. Bearing in mind issues of comparability discussed in Chapter 4, it may have been more appropriate to assess English and Afrikaans subjects separately. Thirdly, little information was provided about how the criteria (grammar and vocabulary, comprehension, fluency and accent) were assessed, and little information was provided about the Foreign Language Memory test (other than that KR-21 values of 0,74 and 0,80 were obtained on samples of 154 Afrikaans speaking and 200 English speaking subjects respectively). In fact, reliability data were not provided for the IMAT, the TSRC or the FLM for the subjects of this particular study.

Fourthly, little information was provided about the content of the French course which the subjects underwent before the criterion data were collected. Finally, Gieseke provides no rationale for her assumption that performance on the TSRCT could be "taken as an indication of ability to cope with similar data in a new language".

Hall (1978) conducted a study as a pilot project to determine whether the regular counselling service at the NIPR could be successfully adapted for use with Black matriculants. The sample comprised volunteer students from Soweto who were attending extra lessons for their matric through the South African Institute of Race Relations. Although tests used in the study were primarily used for the purposes of vocational counselling (i.e as a means of establishing areas of relative strength of students), the results of the sample on the IMAT were compared with the results of a sample of first year university students tested at Fort Hare sample (Visser, 1987). It was found that the mean score of the Fort Hare sample (16,12) was significantly higher than the NIPR test sample (12,78). This result was not altogether unexpected because the population attending university is selected on the basis of academic ability, whereas the other sample was not screened as regards academic ability. The sample used in this study would perhaps not be representative of Soweto matriculants, since, because they had volunteered for both extra matric tuition, as well as the vocational guidance programme, they are likely to have been more motivated and conscientious than the average student.

Hall (1980) conducted a study to investigate the reasons for a high attrition rate on a nursing diploma course. It was hoped that the study would identify factors which would predict successful achievement of the nursing qualification and satisfactory performance in the nursing field. As part of the study a battery of tests including the IMAT were selected which were believed to be relevant to the skills required in nursing. Results on these tests were correlated with the results on College tests and Nursing Council examinations to establish whether they were useful predictors and to determine possible cut-off points. Furthermore the results of those who resigned and those who persisted on the course were compared. It was found that the results of the IMAT and other test of cognitive ability did not prove a useful predictor of success or persistence. Although some trainee nurses had indicated that the academic demands of the course were too great, the study identified a number of other more likely non-cognitive causes for the high drop-out rate. Internal consistency figures (calculated using the corrected KR-21 formula) for the trainee nursing sample were quite reasonable at 0,78.

Halstead (1985) conducted a study to validate three tests of arithmetical ability, namely: the standard level Arithmetic Reasoning Test (ART); the high level Estimations Test (ET) and the Embedded Problems Test (EPT). The IMAT was used as one of several cross-validation instruments. In all three samples (see Tables 1.2 and 1.3) the IMAT was found to correlate moderately at the 1 percent level with all three tests. Somewhat higher correlations were recorded between the three arithmetical reasoning tests and other

tests of numerical aptitude (i.e the Intermediate Arithmetical Problems Test and the Computation Test), and low or non significant correlations were recorded with spatial or verbal ability tests. The moderate correlations between the IMAT and the three tests of arithmetical reasoning ability may be attributed to the fact that the IMAT has a number of arithmetic items included (Wilcocks, 1973). It was unfortunate that test means, standard deviations and reliabilities were not reported for the three samples on which the IMAT was used.

Holburn (unpublished manuscript) conducted a study to evaluate the item bias in several tests including the IMAT. Details of the samples used in this study are provided in Tables 1.3 and 1.4, and details of the findings of the study have already been covered in Chapter 5. Holburn comments that South African literature on testing generally suggest that the "highest mean scores (on psychometric tests) are those of Whites, followed by Asian and Coloureds and then Blacks" (unpublished manuscript, brackets added). The findings of this study are shown to support this general trend,except for the fact that the White mean score in Sample 9 is below the Asian and Coloured mean scores in Samples 6 and 8 respectively. Holburn suggests that this result reflects idiosyncrasies in the sample tested (which is not elaborated) and should not be taken as representative of the population in general.It may be noted that results such as these, supporting group differences in test performance, are the reason for recommendations that fairness policies be implemented in organisations where selection testing occurs across groups (T.R.Taylor, 1987; Holburn, 1990).

With regard to the reliabilities of the test samples, Holburn (unpublished manuscript) comments that these are acceptable in all except the Black Samples 5 and 7, and the Asian Sample 6. The low reliabilities in these groups are attributed to small sample sizes.

Latti (1978) conducted a study to evaluate the prevailing selection procedures for clerks, tellers and typists applying for employment at a particular building society. The IMAT was one of the tests being evaluated since it was part of the prevailing selection battery. Details of the samples involved are provided in Table 1.4. The criterion in the study took the form of a standard evaluation involving a five-point rating of employees: attitudes to others; cooperation; appearance; approach to work; ability to learn; accuracy; speed and concentration (the questionnaire used is provided in an appendix in the study). Criterion assessments were conducted after three months of service and, where possible, after twelve months of service. Amongst all three employee types it was found that significant (at the 5 percent level) but low correlations occurred between the IMAT and the performance criteria after 3 months (see Table 2). Interestingly, the IMAT did not correlate with the work performance criteria after 12 months. Similar low correlations between other tests in the selection battery and the performance criteria led Latti(1978) to question the usefulness of the criteria used.

There are several possible problem areas concerning this study. Firstly, English and Afrikaans samples are combined. It may have been more appropriate to separate these bearing in mind issues of comparability (see Chapter 4). Secondly, while Samples 1 and 2 were relatively large to begin with, they were much diminished after 3 and 12 months when the criterion data was obtained, which once again lays the study open to the problems of small samples. Thirdly, with regard to the obtaining of performance criteria, no mention was made of any attempt to train the raters or standardise their evaluations, which could have resulted in a significant degree of error variance between raters.

Lewis (1980) conducted a study to validate a test battery for the selection of railway chargehands. The IMAT was included in the battery since, as it is a test of general reasoning ability, it was believed to provide a measure of "general trainability". Details of the sample are provided in Table 1.5, which included both Afrikaans and English speaking Whites. A wide range of performance measures were devised as criteria, comprising: verbal communication, activity planning, written communications, information recording, financial awareness, supervisory ability, safety, co-operativeness, leadership, initiative, learning ability, originality, theoretical knowledge, general efficiency and a global performance rating. The employees supervisors were required to evaluate the performance of the employees on a behaviourally anchored seven-point scale, ranging from poor to excellent. Significant (at the 5 percent level) but low correlations were found between the IMAT and three of the criteria, namely verbal communications (0,28), written communications (0,33) and activity planning (0,25). Very low and non-significant relationships were found to exist between IMAT performance and the remaining criteria including a global measure of performance. The internal consistency of responses of the subjects on the IMAT was acceptably high at 0,83 (KR-21 with Tucker correction).

Problems concerning this study are similar to those recorded for Latti (1978) above. Firstly, English and Afrikaans subjects were analysed together, and it may have been more appropriate to separate these bearing in mind issues of comparability. Secondly, with regard to the obtaining of performance criteria, little mention is made of attempts to standardise their evaluations.

Based on the hypothesis that people from different groups have different skills concerning their ability to process various perceptual stimuli, Poortinga (1971) conducted a series of experiments to compare the relative performances of a sample of Black and White undergraduate students on various tests of perceptual ability. In one of the later experiments, the performance of the samples on three tests (including the IMAT) was compared, and their relationship with various tests of perceptual ability was computed. For each of the three tests, the means, standard deviations and odd-even split-half reliabilities were established. On the performance on the IMAT it was again found that there was a large and significant difference in performance between the Black and the White samples, the aggregate Black mean being 12,3, and the aggregate White mean being 23,2. It was found that none of the correlations between the IMAT and the tests of perceptual ability were particularly high.

The major problem of this study concerns the very small sample sizes used (N = 20). It is quite likely that the small sample sizes contributed to the unacceptably low internal consistency figures computed for the Black female undergraduates and the White male undergraduates. Furthermore, it would have been preferable to use the Kuder and Richardson reliability formula (Kuder and Richardson, 1937) rather than the split-half reliability formula (see Section 3.2.3).

Spence (1982) conducted a study to determine a valid procedure to aid in the selection of black guidance teachers for higher-primary schools. The aim of the study was to determine whether certain attributes common to a group of current guidance teachers (cognitive functioning, vocational interest, vocational needs and personality) were significantly different from those of a matched group of non-guidance teachers.It was hoped that the emergence of a clear profile for a Black guidance teachers would assist in the selection and training of future guidance teachers. One of the problems faced by the researcher was finding instruments that could be validly used to assess the attributes of a group of "relatively unsophisticated Black teachers" with a mean level of education of Standard 8 plus a teaching qualification. Spence (1982) noted that previous research had been conducted suggesting the suitability to the IMAT for Black matriculants (Hall, 1978; 1980) and for Black first year students (Visser, 1978, and Erwee, 1981). In an attempt to assess whether the IMAT would be suitable for use with Black guidance teachers, Spence(1982) hypothesised that the level of education of black guidance teachers (Standard 8 plus a teaching qualification) could be regarded as functionally equivalent to a matriculation education. To test this hypothesis Spence compared the performance of a group of Black primary school teachers on the IMAT with that of a group of Black first year university students who had previously been tested (Visser, (1978). Biographical and test distribution details of these two samples are provided in Table 1.4. The mean score of the teacher sample was 10,42 (out of a possible 30), with a standard deviation of 3,96 indicating that there was relatively little spread of the scores, whereas the comparison group of university students showed a mean of 16,57 with a greater spread of score shown by a standard deviation of 4.81. This difference is highly significant, and on the basis of this Spence (1982) concluded that the IMAT would not be suitable for use amongst Black guidance teachers (ie. with a Standard 8 education plus teaching certificate).Spence (1982) reports certain administrative difficulties concerning the administration of the IMAT in the study (for example insufficient invigilators) which may have accounted for the relatively low test scores, but states that this is unlikely to explain the extent of the difference in test score means between the two groups. However, the expectation that Standard 8 guidance teachers with a diploma should perform at the same level as a selected group of first year university students is questionable.

Taylor (1983) conducted research to determine a valid procedure in the identification of potential Engineering Technicians from a sample of Black students (see Table 1.6 for biographical and psychometric details). On the basis of previous research findings, Taylor hypothesised that various measures of intellectual factors, personality factors and scholastic achievements would significantly contribute towards success in the Technikon course in Engineering. As a measure of verbal reasoning ability, the IMAT was included in the test battery to aid in the assessment of "intellectual factors".

As part of the study, Taylor (1983) compared the mean scores of the students who completed their T1 examinations, with those who did not see their course through to the T1 level (drop-outs). It was found that those who dropped out obtained significantly lower mean scores on the IMAT (mean = 16,0) compared to those who completed their first semester examinations (mean = 19,1). The author points out that when interpreting these differences, it should be borne in mind that the T1 drop-out group would include some students who may have been capable of passing T1, but may have been forced to drop out for reasons other than not coping.

The IMAT was found to correlate quite highly with the other tests used in the study. Particularly notable were the correlations with the Intermediate Arithmetical Problems test ($r=0,62$, with a 95 percent confidence interval of $0,43 \leqslant r \leqslant 0,72$), the Standard Level Arithmetic Reasoning Test ($r=0,56$, with a 95 percent confidence interval of $0,42 \leqslant r \leqslant 0,63$) and with the Technical Reading Comprehension Test ($r=0.54$, with a 95 percent confidence interval of $0,41 \leqslant r \leqslant 0,62$). The high correlation with the tests of numerical ability are consistent with the findings of Halstead (1985), and may be explained by the fact that a number of IMAT items have an arithmetical content.

Taylor (1983) suggests that the generally high intercorrelations between the various tests suggest that they are measuring one, or a combination, of the so called "g" factor, plus a speed factor, and a motivational attitude to testing. It was unusual that no significant correlations were found between psychometric tests and the school results of the student sample, and the author suggests that this may indicate that application rather than ability is of primary importance for good scholastic achievement of this target group. Correlations were also computed between the Pre-Tech bridging course results and the IMAT results. The correlation between IMAT performance and averages taken across Pre-Tech Mathematics, Communications, Chemistry, Applied Science and Drawing was 0,45 (with a 95 confidence interval of $0,26 \leqslant r \leqslant 0,60$). This would suggest that the IMAT, together other selected tests, has the potential to predict Pre-Tech subjects fairly efficiently. It was also found that school marks, particularly in science and mathematics were very effective predictors of academic success in the engineering deploma.

Concerning this study it may be commented that the Black sample comprised a number of ethnic groups, which were not compared to check for test score differences, although the sample sizes may have been rather small.

Taylor et al (1982) carried out a study concerned with the programming of the English and Afrikaans versions of the IMAT on the mainframe system PLATO. Part of the study involved the validation of the computerised English and Afrikaans versions of the IMAT against their paper-and-pencil counterparts. For this purpose a number of samples were obtained from a population of White job applicants for clerical and technical positions with Eskom (see Tables 1.6 and 1.7). The standard Eskom selection battery for applicants for technical positions comprises English or Afrikaans versions of the IMAT; Technical Reading Comprehension test; Blox and Mechanical Comprehension test, and the battery for clerical applicants comprises English and Afrikaans versions of the IMAT; Spot-the-Error test; Arithmetic Problems test and Intermediate Reading Comprehension test.

Existing records of recent applicant results on the paper-and-pencil format batteries were obtained from Eskom records. The computerised English and Afrikaans versions of the IMAT were then incorporated into the same batteries and were then used for the next batch of job applicants, and the results compared with those of the existing records.

The obtained results showed that computerisation has little effect on the means, standard deviations and reliability estimates of both English and Afrikaans versions of the IMAT. For both versions and formats of the IMAT, KR-20 reliabilities in the low 0,80 were recorded, which are quite acceptable. In the English version of the IMAT the mean score difference was only 0,24 across test formats (see Samples 1A to 2B in Table 1.6).Inspection of the results across English and Afrikaans samples revealed differences that were found to be significant despite the educational levels of the samples being similar. The mean score differences across English and Afrikaans samples were 2,37 for the computerised tests and 2,50 for the paper-and-pencil Tests. A number of possible reasons for these significant mean score differences were hypothesised and investigated. On the hypothesis that English testees may have resorted more to guessing tactics (which on the IMAT is advantageous since there is no correction for guessing), the ratio of items answered correctly to items attempted was computed. Results showed no evidence to support the hypothesis. Next, on the hypothesis that English and Afrikaans testees may perform differently on different types of items, mean score differences were compared across the 12 verbal items of the IMAT (analogies and logical problems), and across the 18 non-verbal items (series and codes). The results are shown in Table 1.7, Samples 3A and 3B. Comparable differences were found to exist for both sets of items. For the non-verbal items the differences were found to exist for both sets of items. For the verbal items the difference is reported to be 9 percent of the maximum possible score. Hence it is concluded that the differences in performance between language groups cannot be attributed to differential success rates on verbal and non-verbal items.

Next, on the hypothesis that English and Afrikaans job applicants for technical and clerical positions may have performed differently on the verbal and non-verbal items, a similar analysis to the one described above was computed using these variables. Results are shown in Table 1.7, Samples 4A to 5B. In all categories, technical applicants were found to outperform clerical applicants, even on verbal items where clerical applicants may have been expected to be at an advantage. A particularly noticeable mean score difference of 3,16 was found between the English technical and clerical samples. Taylor et al (1982) attribute the high mean score of the English technical group at least in part to the presence of a substantial number of immigrants with technical training. They postulate that the technical training in Britain and Europe may be of a higher standard than local training. It is also possible that recruitment drives from this country may be attracting a higher quality of personnel from overseas.

A final aspect of this study concerned establishing whether computerisation affects correlations between English and Afrikaans versions of the IMAT with other tests. It was hypothesised that the correlations of the computerised versions of the IMAT would be of much the same magnitude as the paper-and-pencil counterparts. The respective correlations of the computerised versus paper-and-pencil, English and Afrikaans versions of the IMAT with the other tests in the Eskom selection batteries may be found in Table 3 (the tests may be found in alphabetical order and the study reference number is 15). Taylor et al (1982) comment that comparison of the correlation variances between the respective English/Afrikaans paper-and-pencil versus computerised versions of the IMAT, and other tests, indicate that computerisation does not appear to add any significant source of variation.

It may be noted that relatively high correlations were recorded between both formats and versions of the IMAT, and all the other tests in the technical selection battery (ranging from 0,40 to 0,65 in magnitude).

Excluding the Blox test which is a test of spatial perception, this could possibly be due the fact that the other tests in the battery require analytical and verbal analyses requiring ability to apply learned knowledge to the solution of problems (i.e similar analytical requirements to the items of the IMAT).It is possible that the relatively high correlation between the IMAT and the Blox test can be attributed to the so - called "G" factor (Anastasi,1982),since the IMAT items were not constructed to assess spatial ability (See Wilcocks,1973). It is interesting to note that Werbeloff and Taylor (1982) also recorded relatively high correlations between the IMAT and tests of spatial perception for their Black test samples.

In the clerical battery, relatively high correlations were obtained between the IMAT and the Intermediate Arithmetical Problems and Intermediate Reading Comprehension tests. These may be expected since the IMAT contains items requiring numerical and verbal analyses. Low correlations between the IMAT and the Intermediate Level Spot-the-Error (Accuracy) tests are not particularly surprising since the latter is a measure of perceptual ability. This study was one of the more rigorous and methodically thorough of the studies reviewed. There was a relatively large study sample (although not of the magnitude necessary to cater for the seven sources of variance listed by Hunter et al, 1982). Hypotheses were logically formulated, and details of the study samples, methods and test results clearly presented.

Based on a proposed need to initiate a counselling service for Black matriculants (Hall, 1978), Visser (1978) undertook to validate a number of NIPR tests (including the IMAT) for use in counselling or selecting Black students. It was decided that the criterion for the validation would be success at university. Details of the two samples used in the study are presented in Table 1.9. While Sample 1 initially comprised 218 Black first year students at Fort Hare University and Sample 2 initially comprised 373, somewhat less than these numbers of students came to the testing after the preliminary, biographical data collecting session. Concerning Sample 1, which was tested in 1977, it was found that the Intermediate Battery tests showed a better reliability and better spread of scores than the High Level Battery. As a whole the High Level Battery tended to show a positively skewed score distribution showing that the majority of testees found the items too difficult. A similar pattern of scores was found concerning Sample 2, which was tested in 1978. Intercorrelations computed between the results of the two samples on the various tests revealed that the IMAT tended to correlate relatively highly with the Ravens Progressive Matrices test ($r = 0,54$ and 0,61 respectively), the Intermediate Computation test ($r = 0,53$ and 0,58 respectively), and to some extent with the Intermediate Reading Comprehension test ($r = 0,28$ and 0,40 respectively). This tends to support the findings of Epstein (1983), Erwee (1981) and Hall (1978) that IMAT items sample a wide domain of aspects of formal education, including numerical reasoning, non-verbal conceptual reasoning and reading comprehension.

Sample 1 was evaluated with regard to establishing the predictive utility of the tests. For this purpose the University of Fort Hare supplied half-year marks and end-of-year marks which served as criterion data. The major problem with this exercise was the small numbers in the samples (ranging from 5 to 17), the reasons for which are left vaguely explained. Therefore despite the fact that IMAT results were found to correlate significantly with Private Law ($r = 0,54$), Statistical Methods ($r = 0,90$) and English ($r = 0,63$), these figures should only be considered as tentative, requiring further validation in larger samples. Concerning the correlation between tests and the end of year results, Visser (1978) reports that incomplete lists of the final year results led to numerous cases being lost. Nonetheless, test results were correlated with available examination results data, and it was found that other than with Chemistry ($r = 0,39$: with a 95 percent confidence interval of $0,13 \leqslant r \leqslant 0,42$) the IMAT did not correlate with the year-end examination results. Unfortunately the exercise to establish the predictive utility of the tests was not repeated with the 1978 sample,and so no comparison was possible between years. Visser (1978) identified a particular problem concerning the calculation of a single criterion score which was comparable for all students, and which represented the quality of academic success for the year.

It was a disappointing aspect of this study that the predictive utility of the IMAT could only be evaluated against various course results with such small samples. It was not sufficiently reported how a single criterion figure was arrived at with regard to the measure of academic success. The mean score, deviation and internal consistency data of the IMAT were relatively constant across the two years of testing (see Table 1.9).

Visser (1980) conducted a study to investigate the implications of electroencephalographic (EEG) assessments in the vocational guidance of NIPR clients with suspected brain abnormalities. The sample comprised clients who had come to the NIPR for vocational guidance counselling between 1965 and 1978, and were referred for EEG on the counsellors initiative (see Table 1.10 for details). Indications for referral were lack of concentration, restlessness, behaviour problems, academic underachievement or vocational instability,and particularly biographical data showing evidence of some form of head injury. Although there was no stated hypothesis, the test results of the EEG sample as a whole were compared with a group of South African Railways male clerical staff tested in 1964 and 1965. It was found that the mean score of the EEG group (15,01) was slightly less than that of the comparison group (16,43), but a significance test indicated that this difference was non-significant.

In a second phase of this study, an attempt was made to differentiate the EEG sample into one of three groups, according to the severity and nature of their complaints (see Table 1.10 for details). In view of a generally observable discrepancy between the IMAT test results of the EEG group and the population group, it was decided to ascertain whether any one of the three groups was responsible for the lower scores. Application of the Kruskal-Wallis one-way analysis of variance (Siegel, 1956) to the mean scores of the different groups (see Table 1.10) showed that these did not differ significantly from each other. Visser then provided a number of case histories of the career counselling and career paths of some of the EEG sample, and indicated that appropriate referral on the basis of abnormal EEG readings may affect the counsellees career development positively.

There are several points of issue concerning this study. Firstly, no formal hypotheses were set out as the basis of the study. Secondly, subjects were selected over a wide time span, laying test results open to time related sources of variance. Thirdly, the group chosen for comparison against the EEG group was somewhat puzzling, primarily because it was specifically a sample of male, railways clerical staff. It may have been more appropriate to match the EEG sample against a sample of other vocational guidance clients. Fourthly, internal consistency figures are not provided for the various research samples. Fifthly, the sample sizes of the three sub-groups of the EEG sample were particularly small (N = 26, 19 and 23 respectively), and results from such small samples could not be considered generalisable in any sense. Finally, little information is provided as to exactly how the EEG group was classified into its sub groups. Visser notes that an attempt was made to be consistent in classifying the cases, but that information was rarely obtained from medical sources, and therefore discrepancies could easily have arisen in assessing their significance and severity.

Finally, Werbeloff and Taylor (1982) conducted a study concerned with the history, construction and validation of the High Level Figure Classification (HLFCT). The HLFCT is a test of non-verbal reasoning ability, and was developed to meet a need to assess the cognitive abilities of people in the Std. 8 to 10

education range (Werbeloff and Taylor, 1982). Validation was carried out on eight samples of Asian or Black subjects. On two of these samples performance on the HLFCT was correlated with that on various other tests including the IMAT. The first of these was a sample of Black male applicants applying for training as technicians for electronic and electromechanical equipment, and the second was a group of first year students of Mabopani East Technikon (see Table 1.10). Psychometric statistics on the IMAT, and information on where the applicant technicians were applying from, was not provided.

In both samples, the IMAT was found to correlate significantly (at the 5 percent level) with all of the other tests used in the study, namely the Raven's Progressive Matrices, the Technical Reading Comprehension Test, the Standard Level Arithmetic Reasoning Test, the Estimation test, the Blox test, the "F" test and the "H" test. As was expected the IMAT correlated significantly with the HLFCT (in Sample 1:r = 0,60: with a 95 percent confidence interval of $0,51 \leqslant r \leqslant 0,67$; and in Sample 2: r = 0,48, with a 95 percent confidence interval of $0,38 \leqslant r \leqslant 0,56$), since both are tests of reasoning ability,although differences in performance are to be expected since the IMAT is a test of verbal reasoning while the HLFCT is a test of non-verbal reasoning. Once again the moderate correlations of the IMAT with the other tests in the study support the findings of Epstein (1983), Hall (1978) and Halstead (1985) that the IMAT samples a wide domain of abilities including numerical reasoning, non-verbal reasoning, inductive reasoning and reading comprehension. While the correlations between the IMAT and the tests of spatial perception (Blox, "F" and "H") are generally lower than the other tests,Werbeloff and Taylor (1982) attribute the recorded relationship to the possibility that the IMAT may sample elements of spatial ability in a Black sample.However, another possibility that has already been suggested in connection with the study of Taylor et al (1982),is that the relationship could be attributed to the so - called "G" factor.

In Sample 1, a second component to the research involved correlating performance on the tests (including the IMAT) with criterion data in the form of end-of-year examination results (electronics, applied mechanics, typewriting, adding machine, electronics self study). No information was provided on the content domain of the courses or examinations. The IMAT was found to correlate significantly (at the 5 percent level) with all of the criteria, namely electronics, applied mechanics, typewriting, adding machine and self study. Respective correlations and 95 percent confidence intervals for these criteria were $r=0,49$ $(0,30 \leqslant r \leqslant 0,65)$; $r=0,32$ $(0,11 \leqslant r \leqslant 0,50)$; $r=0,24$ $(0,0 \leqslant r \leqslant 0,46)$; $r=0,29$ $(0,03 \leqslant r \leqslant 0,43)$ and $r=0,50$ $(0,30 \leqslant r \leqslant 0,65)$. In all cases the correlations between the criteria and the HLFCT were slightly higher than those betweem tje criteria and the IMAT, but the HLFCT and the IMAT correlated higher with the criteria then the other tests in the study did. Werbeloff and Taylor (1982) conclude that the HLFCT appears to perform better than the IMAT as a predictor of course success in the case of the selected group of Black male technicians, but that the tests are by no means interchangeable since they are only moderately correlated, and part of their predictive power is likely to come from their respective verbal or non-verbal characteristics. They therefore propose that both tests are used in the selection of Black technicians.

In addition to the disadvantages of small sample studies there are several points of issue concerning this study. Firstly, incomplete information is provided on the test samples (see Tables 1.10 and 1.11), and the IMAT test performance characteristics of the two samples are unfortunately not provided. Secondly, as noted above, no information is provided on the content domain of the courses or examinations.

Having thus reviewed the different studies on the IMAT, at this stage it may be useful to reconsider the effect of restriction of range on correlation coefficients. In Section 3.2.4.2 it was noted that the size of a correlation coefficient depends partially upon the heterogeneity of the population from which the data is obtained. In a homogeneous population the range of ability is likely to be restricted. The narrower the range, the smaller the correlation coefficient tends to be (Anastasi, 1982).

Perusal of Tables 1.1 to 1.11 indicates that most of the studies report restriction in the samples in terms of education, age and ethnic group. This suggests that correlations obtained with these samples on the criteria reported would be lower than in samples with wider ranges in terms of education, age, and racial or ethnic groups. Bearing in mind that the samples in all of the studies reviewed were subject to some restriction of range, it may be useful to consider patterns in the IMAT standard deviations of the samples when evaluating the correlations obtained by these samples.

The range in IMAT standard deviations across studies was 3,32 to 6,40. The mean standard deviation was 4,68. This indicates that in relation to all of the samples reviewed, samples with standard deviations closer to 3,32 (i.e. with a relatively narrow range of scores) would tend to obtain lower correlation coefficients than samples with standard deviations closer to 6,40 (ie. with a relatively wider range of scores).Notably low standard deviations were recorded in Study 8 for Sample 5 (3,72), Sample 6 (3,62) and Sample 7 (3,62) ; in Study 12 for Sample 3 (3,84) and Sample 4 (3,32), and in Study 13 for Sample 1 (3,96).Had all of these studies involved correlation of test data with criteria, these correlations may have been expected to be slightly depressed because of the restriction in score range. However of these studies, only Study 12 involved correlation, and this study only correlated the IMAT with Blox and Advanced Ravens test. It was found that Samples 3 and 4 (combined, both with notably low standard deviations), obtained a correlation of 0,36 with the Blox test, and correlation of 0,29 with Advanced Ravens test. In both instances these correlations tended to be lower than corresponding correlations computed in other studies (see Tables 3.1(b) and 3.4(a), suggesting the expectation of finding lower correlations in retricted score ranges.

Conversely, notably high standards deviations were recorded in the Study 1 for its only sample (6,40), in Study 8 for Sample 9 (6,22), and in Study 12 for Sample 1 (6,00). Once again not all of these studies involved the correlation of data. Where correlations were made, however, these tended to be relatively high. The sample of Study 1 obtained correlations between 0,41 and 0,74 (mean 0,61) on all seven of its criteria (see Tables 2.1, 3.1(a) and 3.5(a)). Similarly, relatively high correlations were recorded for Sample 1 of Study 12. These correlations were respectively 0,59 with the Blox test and 0,73 with Advanced Ravens test, tending to be higher than corresponding correlations computed in other studies (see Table 3.1(b) and 3.4(a)). Such relatively high correlations suggested that they might have been affected by the relatively wide score range.

In conclusion, where given correlations are being compared across samples, it may be useful to evaluate the range of standard deviations across samples. The more that standard deviations differ from the mean standard deviation, the more that the correlations obtained from these sampels may be expected to differ from the "true" correlation.

CHAPTER 7

CONCLUSIONS AND RECOMMENDATIONS

In Section 1.1 it was mentioned that despite the fairy important role of the IMAT, particularly for apprentice selection in industry, there has been no comprehensive study to review the applications of this test. Bearing in mind that the test has been in existence for three decades, when the current study was originally planned, it was envisaged that in the course of research on the IMAT it would be possible to find a sufficient number of studies to identify common areas of research, for example the correlation of the IMAT with T1 examination results, or with other specific criteria. Such replication of studies are necessary to confirm trends in the research, particularly with the aid of the statistical technique of meta-analysis (see Section 3.1.1).

The review of studies involving the IMAT reveals few common areas of research and little consistency in reporting standards.The establishment of consistent trends in research findings, with or without the aid of meta-analysis, requires some breadth of data from parallel set of research circumstances (Hunter et al, 1982; Tukey, 1969). In line with the proposals of Tukey (1969) covered in Section 1.2, this would indicate a need for further research on the IMAT to build on the hypotheses and findings of studies that have already been conducted, so as to create a more consistent and richer body of data from which to work. It is hoped that the current integrative review will serve as a starting point for additional research by highlighting research studies and findings on the IMAT to date, and indicating the minimum psychometric reporting criteria for future work in this direction (this will be provided at the end of this chapter).

While not stipulating a minimum number of studies with common criteria in order to use the meta-analysis technique, Hunter et al (1982) use a minimum number of six studies in their own application of the technique. Since it was not possible to trace even two validity studies on the IMAT with common criteria, it is clearly not possible to use the meta-analysis technique to generalise the validity of the IMAT. It is thus necessary to review the studies individually, and then attempt a qualitative integration of research across studies. When interpreting validity findings, there is a temptation to regard given correlation data as sufficient evidence for the existence of particular relationship between given variables. Because of the potential error variance which will affect research using "small samples", or samples with less than a thousand subjects, Hunter et al (1982) primarily advocate the use of confidence intervals (see Section 3.1.2.) which "...give a correct picture of the extent of uncertainty that surrounds results computed from small sample studies.. " (Hunter et al, 1982,p.24). It is notable that confidence intervals for correlations were not provided in any of the studies reviewed. Instead authors appear to have regarded

one-off correlations studies as sufficient proof of statistical significance for the acceptability and generalisability of their findings without taking into account the extent of possible confidence range due to sources of error variance. The work of Schmidt, Hunter and their colleagues that has been reviewed in Section 4.2.7. above has clearly shown that this approach to the interpretation of individual studies is problematic, and is likely to lead to errors in conclusions. Bearing in mind the limitations of small sample research, and the fact that there has been relatively little consistency in the research criteria against which the IMAT has been evaluated, an attempt will now be made to provide a qualitative integration of the nomological network of meaning of the findings that have been summarised in this study. Although the manual for the IMAT (Wilcocks, 1973) does not stipulate a specific original purpose for which the test was constructed, it may be useful to evaluate the IMAT as a "verbally based omnibus measure of mental ability" (Taylor,Werbeloff and Ebertsohn, 1982,p.10), or as "a measure of general verbal reasoning ability (Wilcocks, 1973).

Looking at Tables 1.1 to 1.11, it can be seen that the IMAT has been used as a measure of cognitive functioning against which different groups have been compared (Epstein, 1983; Erwee, 1981; Hall, 1978; Holburn ,unpublished manuscript; Poortinga, 1971; Spence, 1982; Visser, 1980), as a predictor (Christierson, 1977; Gieseke, 1970; Hall, 1980; Latti, 1978; Taylor, 1983; Visser, 1978), and as a validation instrument for other tests (notably Halstead, 1985; and Werbeloff and Taylor,1982). Findings from these three different kinds of studies will now be summarised, and general issues raised.

It was observed that in the studies evaluating the IMAT performance across different groups, subjects were differentiated on the basis of racial groups rather than ethnic group. For example, English and Afrikaans speaking subjects were taken together as a "White" sample in many of the studies reviewed eg. Christierson (1977), Gieseke (1970), Hall (1980), Visser (1980). Yet in one study where English and Afrikaans subjects were differentiated (Taylor et al, 1982), there was a marked difference in group means on the IMAT, with English speaking subjects tending to perform better. This is consistent with a marked mean score difference that occurred in the original standardisation samples between English speaking subjects (Standard 10 mean = 20,09 ; Standard 7-9 mean = 16,43) and Afrikaans speaking subjects (Standard 10 mean = 18,10 ; Standard 7-9 mean = 13,89) - (Wilcocks, 1973). Bearing in mind such mean score differences, researchers who mix Afrikaans and English speaking subjects may unwittingly be adding a source of variance to their results and should check to see if language is a moderator variable. It is possible that similar differences exist between the different Black ethnic sub groups, and it is therefore recommended that in any future research, ethnic as opposed to racial groups be compared (see Chapter 5). This would provide greater heuristic insight into the numerous factors that can cause differences in mean scores across groups.

Consideration of the test score distributions of studies evaluating the IMAT performances across different racial groups tends to confirm the assertion of Holburn (unpublished manuscript) that "...Generally in South African literature we find the highest mean score are those of whites, followed by Asians and coloureds, then blacks". Although the comparability of different test scores across different racial groups is suspect (see Chapter 4), the studies reviewed have clarified the population groups on which it is best to apply the IMAT. The IMAT was originally standardised on White English and Afrikaans speaking, male, South African Railways clerical staff with 9 to 12 years of education (Wilcocks,1973). However, there is research to suggest that the High Level Mental Alertness Test shows better reliability and yields a more normal distribution than the IMAT for White higher grade matric students. Conversely, Visser (1978) reports that the IMAT yields a more normal distribution and a better reliability as far as Black first-year university students are concerned, and is thus a more appropriate test for Blacks of that level. This finding has been supported by Erwee (1981) and Hall (1978, 1980) who have demonstrated that the IMAT may be used successfully with Black matriculants. Research conducted by Spence (1982) suggests that the IMAT is not suitable for use with Black Standard 8 students.

Moving on now to consider the studies using the IMAT as a predictive instrument, Christierson (1977) showed that the IMAT had a fairly high positive correlation with the criterion of training success on the Civil Engineering Industry Training Board's basic training course in instructional techniques. Gieseke (1970) showed that the IMAT had fairly high positive correlation with criteria associated with the ability of Naval recruits to learn a new language, namely French vocabulary and grammar, French comprehension and French fluency. The results of a study by Hall (1980) indicated that performance on the IMAT was not a useful predictor of persistence or success on the nursing diploma course, where the emphasis may have been more on rote learning. The results of the study by Latti (1978) indicated that the IMAT did not correlate particularly highly with performance of clerks, tellers and typists at a particular building society, where routine activity would have been more important than general reasoning ability. Similarly Lewis (1980) found that the IMAT was not a useful instrument in the prediction of successful railway chargehands. Although not clearly indicated,there was probably little emphasis on analytical and conceptual problem solving and more emphasis on routine in the work performance being predicted. Taylor (1983) found that the performance of a sample of Black engineering technician students on the IMAT correlated positively and quite highly with Pre-Tech bridging course subjects, particularly Mathematics and Communication. Visser (1978) found that the performance of a sample of Black first year university students on the IMAT correlated positively and quite highly with a number of university courses such as Private Law, Statistical Methods and English. These two findings emphasise the usefulness of the IMAT in predicting academic performance.

Moving now to consider the studies involving the correlation of the IMAT with other aptitude or ability tests indicates that the IMAT tends to correlate consistently significantly and highly with particular types of tests. Tables 3.1(a) indicates that across studies the IMAT correlated between 0,48 and 0,75($\bar{X}$ =0,63) with the Intermediate Arithmetical Problems Test. Similarly, it can be seen that with one exception, the IMAT correlated between 0,42 and 0,57($\bar{X}$=0,50) with the Standard Level Arithmetical Reasoning Test. A slightly lower, but nonetheless moderate correlation between the IMAT and the Standard Level Arithmetical Reasoning Test was recorded for the 263 Black male technician applicants (r=0,37, with a 95 percent confidence interval of $0,27 \leqslant r \leqslant 0,46$) in Study 18. However,if this is contrasted with the high correlations recorded for 252 Black male technician applicants in Study 2 (r=0,61, with a 95 percent confidence interval of $0,57 \leqslant r \leqslant 0,71$), and for the 149 Black male T1 engineering students at Mongosuthu technikon in Study 14 (r=0,62, with a 95 percent confidence interval of $0,51 \leqslant r \leqslant 0,69$), and for the 363 Black first year university students in Study 16 (r=0,48, with a 95 percent confidence interval of $0,39 \leqslant r \leqslant 0,56$), the "true" correlation between the IMAT and the SLART would appear to be somewhat higher than this exception. Table 3.3(a) indicates that across studies the IMAT correlated relatively highly (between 0,48 and 0,60) for both the Standard and the Higher Level Figure Classification Tests. Table 3.5(a) and 3.5(b) indicates that across studies the IMAT generally correlated between 0,40 and 0,72 with the Technical Reading Comprehension Test. One lower correlation between the IMAT and the Technical Reading Comprehension Test was recorded for the 199 Black T1 students in Study 18. However, once again comparing this study with the findings of the other studies with similar samples, this appears to be the exception rather than the rule.

Table 3.4(a) indicates that across studies the IMAT generally correlated between 0,50 and 0,73 with both the Standard and the Advanced Level Ravens' matrices. There are two exceptions to this trend. The first was recorded between the IMAT and the standard level Ravens' matrices for the 110 Black first year students in study 3 (r=0,33, $0,15 \leqslant r \leqslant 0,48$), and the second was recorded between the IMAT and the Advanced Level Ravens' matrices for 40 White female undergraduate students in study 12 (r=0,29, with a 95 percent confidence interval of $-0,5 \leqslant r \leqslant 0,53$). Concerning the first exception, the correlations recorded between the IMAT and the Standard Level Ravens' matrices for similar samples in Studies 14 and 16 suggest that the finding is likely to be lower than the "true" correlation. Concerning the second exception, the low recorded correlation will almost certainly have been affected by the small sample size (N=40) which is reflected in the wide range of the confidence intervals. Table 3.4(a) indicates that across studies the IMAT generally correlated between 0,40 and 0,69 with the Intermediate Reading Comprehension Test. It is notable that the two exceptions were both recorded in the case of Black first year university students. The first was recorded using Sample 1 and 2 of Study 3 (r=0,26, with a 95 percent confidence interval of $0,07 \leqslant r \leqslant 0,42$) and the second was recorded using Sample 1 of Study 16 (r=0,26, with a 95 percent confidence interval of $0,09 \leqslant r \leqslant 0,41$). However, it is difficult to find a plausible explanation for this other than the effects of error variance, particularly since Sample 1 of Study 16 (also Black first year university students) produced a somewhat higher correlation between the IMAT and the Reading Comprehension Test (r=0,40, with a 95 percent confidence interval of $0,30 \leqslant r \leqslant 0,49$).

The tables of correlation between the IMAT and other tests reveal an unexpected yet consistent correlation between the IMAT and tests of spatial ability, notably with the Blox test (see Tables 3.1(a) and 3.1(b)). Apart from one or two exceptions, correlations between the IMAT and the Blox test can been seen generally to range between 0,38 and 0,62. These correlations occur across samples of White chargehands, clerical workers and technicians, and Black apprentice applicants, T1 students and undergraduate university students. Bearing in mind the range of samples across which the correlation occurs, it appears likely that this phenomenon can be attributed to the so called "G" factor (Anastasi, 1982),since the items do not appear to relate to spatial ability.

As has been stated above, review of studies indicates that the IMAT has been correlated with a wide variety of criteria, but unfortunately there has been relatively little replication of research criteria across studies. Tests that were observed to correlate relatively highly with the IMAT in "one-off" studies include the following (with confidence intervals provided at the 95 percent level):

- Concept Identification Test: $r=0,45$; $0,33 \leqslant r \leqslant 0,56$
- Deductive Reasoning Test (Intermediate): $r=0,50$; $0,35 \leqslant r \leqslant 0,62$
- Figure Analogy Test: $r=0,56$; $0,41 \leqslant r \leqslant 0,65$
- Foreign Language Memory Test: $r=0,59$; $0,32 \leqslant r \leqslant 0,75$

Taken together, the relatively high correlations of the IMAT with the range of tests listed above can be viewed as a rudimentary nomological network (see Section 3.3.4) from which evidence can be gleaned concerning the construct validity of the IMAT. The correlations with criteria listed above suggested that the IMAT samples a relatively wide domain of abilities, particularly arithmetical ability, verbal comprehension ability and inductive reasoning ability. This is consistent with factor analysis studies indicating that the IMAT loads highly on factors of analytical reasoning and conceptual reasoning (Epstein, 1985),and verbal ability and quantitative - analytical ability (Wilcocks, 1973). Together, this "network" lends credibility to the fact that the IMAT measures the construct of "general reasoning ability". It may be noted that the IMAT correlated at a relatively low level with tests requiring manual dexterity (the 0'Conner Test of Finger Dexterity), hand-eye co-ordination (the Poppelreuter) and perceptual speed (the Spot-the-error Test). This is not unexpected since the IMAT is primarily a reasoning test and should not correlate with manual dexterity, co-ordination or perceptual speed.

Overall, the various studies that used the IMAT revealed a number of problem areas. One of the problems encountered was that of small sample sizes, which was particularly noticeable in the studies of Gieseke (1970), Poortinga (1971), and Visser (1978). While it is recognised that obtaining or processing data of the "one thousand or more" subjects recommended by Hunter et al (1982) is unlikely to be feasible, studies employing samples in the vicinity of 20 people are unlikely to provide useful data.

A second problem emerged in the studies assessing the predictive validity of the IMAT, where little or no information was provided on the nature, reliability or validity of the criterion assessment instrument. Furthermore, where raters were involved in the assessment of criteria,little or no mention was made of attempts to train them or ensure that their ratings were reliable and valid.

A third problem encountered in the studies reviewed was that the actual hypothesis being tested was often left unstated and assumed. Where statistical significance is being evaluated, the nature of the hypothesis being tested is of vital importance. For example, in studies providing correlational data that is significant at the 5 percent level, the 5 percent error rate is only guaranteed if the null hypothesis is true. If the null hypothesis is false the error rate can go up to 95 percent (Guilford, 1965).

Following on from the last point, a fourth problem that emerged was simply that useful or important information was not always reported,such as the IMAT means, standard deviations, internal consistency figures, dates when these samples were tested, how the subjects were sampled,and biographical information on the subjects tested. If integrative studies using meta-analytic techniques are ever going to be conducted it is important that essential data is included in the research report. Hunter et al (1982) provide a list of the essential data required for the metra-analysis of correlational studies. These are:

- the sample mean
- the sample deviation
- the internal consistency reliability
- the reliability of the criterion measurement instruments
- information pertaining to the sampling of the subjects
- the entire matrix of zero-order correlations between all variables (Hunter et al (1982) note that the means, standard deviation and reliabilities can easily be appended as extra rows or columns of the matrix).

In conclusion, there is a relative scarcity of published research on the IMAT. There is a need for studies to cover a more consistent research domain, preferably being based on the hypotheses or research findings of previous studies on the IMAT. Attention needs to be given to the sampling of the subjects (see Cole and Means, 1981), as well as the sample size, which should be more than 30 in size, and preferably more than 100. Where it is not possible to work with samples of 1 000 or more suggested by Hunter et al (1982) to eliminate sources of variance, it is useful to provide confidence intervals of correlations rather than the more traditional statistical significance figures - which may lead to errors in interpretation (Hunter et al, 1982). Where studies are reported, it is important that these are comprehensively described with details of hypotheses tested, sampling procedures, dates of assessment, data on the distribution of test scores, and the methods of assessment of the criteria in the study. Rather than conducting research on samples differentiated according to arbitrary racial classification, it is recommended that samples be differentiated on the basis of ethnicity and\or socioeconomic status which is likely to lead to a greater understanding of the nature of group differences, with potential benefits for fair testing in the applied context.

APPENDIX 1.1: NORMS FOR THE IMAT

SAMPLE DESCRIPTION	TESTING LANGUAGE	N	MEANS	SD	RELIABILITY	DATE TESTED
1. Std. 8 - 9 Black males tested for BEATS, Soweto	English	153	11,18	3,82	0,713 (KR-21 + Tucker)	Not stated
2. Std. 8 - 10 Black males tested for BEATS, Soweto	English	253	13,21	4,63	0,798 (Kr-21 + Tucker)	Not stated
3. Std. 10 Black males tested for BEATS, Soweto	English	100	16,30	4,02	0,717 (KR-21 + Tucker)	Not stated
4. St. 10 Black male and female students at Pace College (a commercial school), Soweto	English	125	17,13	3,44	0,630 (KR-21 + Tucker)	1985
5. 1 st year Black male and female students at University of the North (ages 17-25 yrs)	English	132	16,12	4,74	0,806 (KR-21 + Tucker)	1977
6. 1 st year Engineering Technicians of Mangosuthu Technikon (average age 22 years)	English	149	18,50	4,60	0,810 (KR-21 + Tucker)	1980 - 1981
7. Std. 8 - 10 Black male apprentice applicants to the Natal Sugar Industry (ages 18-30 yrs)	English	220	14,34	4,53	0,783 (KR-21 + Tucker)	1981 - 1985
8. Std. 8 - 10 Black male apprentice applicants to the Natal Sugar Industry (ages 18-27 yrs)	English	208	12,27	4,18	0,752 (KR-21 + Tucker)	1988

APPENDIX 1.2: NORMS FOR THE IMAT

SAMPLE DESCRIPTION	TESTING LANGUAGE	N	MEANS	SD	RELIABILITY	DATE TESTED
9. Std. 8 - 10 Coloured apprentices to the Natal Sugar Industry (ages 17-31 yrs)	English	102	17,36	4,88	0,825 (KR-21 + Tucker)	1988
10. Std. 8 - 10 Asian apprentices to the Natal Sugar Industry (ages 18-27 yrs)	English	206	17,47	4,75	0,814 (KR-21 + Tucker)	1988
11. Std. 7 - 10 Asian male apprentices applicants to the Natal Sugar Industry (ages 18-32 yrs)	English	428	16,66	5,03	0,388 (KR-21 + Tucker)	1981 - 1985
12. Std. 8 - 10 White apprentices to the Natal Sugar Industry	English	99	18,75	4,48	0,800 (KR-21 + Tucker)	1988
13. Matriculant White male and female clerical staff at a Building Society (ages 17-35 yrs)	English	103	17,85	4,64	0,807 (KR-21 + Tucker)	1984
14. Non university entrance White matriculant males and females reporting to the NIPR for Vocational guidance	English	182	17,84	4,66	0,806 (KR-21 + Tucker)	1980 - 1983

APPENDIX 1.3: NORMS FOR THE IMAT

SAMPLE DESCRIPTION	TESTING LANGUAGE	N	MEANS	SD	RELIABILITY	DATE TESTED
15. Std. 7 - 9 White male clerical staff employed with South African Railways (mean age 19, 25)	English	205	16,43	4,88	0,751 (KR-21 + Tucker)	1964 - 1965
16. Std. 10 White male clerical staff employed with South African Railways (mean age 18, 77).	English	176	20,09	4,31	0,715 (KR-21 + Tucker)	1964 - 1965
17. Std. 10 White (155 females, 14 males) student teachers at the Johannesburg College of Education (mean age 18,22).	English	169	20,43	3,69	0,607 (KR-21 + Tucker)	1965
18. Std. 7 - 9 White male clerical staff employed with South African Railways (mean age 19, 41).	Afrikaans	803	13,89	4,98	0,761 (KR-21 + Tucker)	1964 - 1965
19. Std. 10 White male clerical staff employed with South African Railways (mean age 19, 50).	Afrikaans	548	18,10	3,52	0,764 (KR-21 + Tucker)	1964 - 1965
20. Non university entrance White males and females reporting to the NIPR for Vocational Guidance.	Afrikaans	132	17,18	?	?	1989

REFERENCES:

Aiken, L.R. (1982) Psychological testing and assessment (4th ed.). Boston: Allyn and Bacon.

American Psychological Association. (1975) Division of Industrial - Organisational Psychology. Principles for the validation and use of personnel selection procedures. Dayton, Ohio: Author.

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

Bennett, G.K. Seashore, H.G. and Wesman, A.G. (1960) Fifth edition manual for the differential aptitude tests, Forms S and T. New York: Psychological Corporation.

Biesheuvel, S. (1943) African intelligence. Johannesburg: South African Institute of Race Relations.

Biesheuvel, S. (1949) Psychological tests and their application to non-European people. In G.B. Jeffrey (Ed.). The yearbook of education. London: Evans.

Biesheuvel, S. (1952) Personnel selection tests for Africans. South African Journal of Science. 49: 3 - 12.

Biesheuvel, S. (1958) Objectives and methods in African psychological research. Journal of Social Psychology. 47: 161 - 168.

Biesheuvel, S. (1960) "Inter-African Psychological Research" : Proceedings during congress, held at the University of Natal in Durban from 13 - 20th July 1960. Proceedings of the South African Psychological Association. No. 9/10. 23 -25.

Biesheuvel, S. (unpublish manuscript) History of the National Institute for Personnel Research.

Biesheuvel, S. (1987) Psychology: science and politics. Theoretical developments and applications in a plural society. South African Journal of Psychology. 17: 1-8.

Boring, E.G. (1950) A history of experimental psychology (Rev. ed). New York: Apleton-Century-Crofts.

Buros, O.K. (Ed.) (1959) Review of the Otis Quick Scoring Mental Ability Tests. Fifth Mental Measurements Yearbook. New Jersey: Gryphon.

Campbell, D.T. and Fiske, D.W. (1959) Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin. 56:81-105.

Christierson, V.A.B. (1982) Selection of trainees for a course in instructional techniques in the field of civil engineering - a validation study. CSIR Special Report, PERS 252 Johannesburg: National Institute for Personnel Research.

Cole, M., and Means, B. (1981) Comparative studies of how people think. Cambridge: Harvard.

Crawford-Nutt, D.H. (1977) The effect of educational level on the test scores of people in South Africa. Psychologia Africana. 17, 49 - 59.

Cronbach, L.J. (1951) Coefficient alpha and the internal structure of tests. Psychometrica. 16, 297-334.

Cronbach, L.J. (1970) Essentials of psychological testing (3rd ed.). New York: Harper and Row.

Cronbach, L.J., and Gleser, C.C.,(1965) Psychological tests and personnel decisions. (2nd Ed.). Urbana: University of Illinois Press.

Donald, C., Veldsman, T., Donald, F., Cook, J., Chemel, C., and Taylor, T. (1990) Turning fairness in selection and placement into practical company policy. S A Journal of Labour Relations. 14, 52 - 77.

Dubois, P.H. (1970) A history of psychological testing. Boston: Allyn and Bacon.

Epstein, B.I. (1985) Factors related to mechanical aptitude in Blacks II: Empirical study. CSIR Special Report, PERS 381. Pretoria: Human Sciences Research Council

Erwee, R. (1981) Cognitive functioning, achievement motivation and vocational preferences of Black university students. Psychologia Africana. 20,29 - 51.

Ferguson, L.W. (1962) The heritage of industrial psychology. Hartford, C T: Finlay.

Gardner, J.M. (1980) Post graduate training for non-whites in clinical and counselling psychology in South Africa: A survey. S A Journal of Psychology. 10,6 - 10.

Ghiselli, E.E. (1966) The validity of occupational aptitude tests. New York: Wiley.

Gieseke, M. (1970) Predicting the ability to learn a foreign language. Psychologia Africana. 13, 218 - 221.

Glass, G.V. (1976) Primary, secondary and meta-analysis of research. Educational Researcher. 5,3-8.

Glass, G.V. (1977) Integrating findings: The meta-analysis of research. Review of Research in Education. 5,351 -379.

Goddard, H.H. (1910) A measuring scale for intelligence. The Training School. 6, 146 - 155.

Goodenough, F.L. (1949) Mental testing: Its history, principles and applications. New York: Rinehart.

Guilford, J.P. (1965) Fundamental statistics in pychology and education. New York: McGraw-Hill.

Guion, R.M. (1965) Personnel testing. New York: McGraw-Hill.

Guion, R.M. (1976) Recruiting , selection and job placement. In M.D. Dunnette (Ed). Handbook of industrial and organisational psychology. Chicago: Rond McNally.

Guion, R.M. (1980) On trinitarian doctrines of validity. Professional Psychology. 11 (3), 385 - 398.

Hall, B.A. (1978) Vocational couselling for Blacks with high school education. CSIR Special Report, PERS 270. Johannesburg: National Institute for Personnel Research.

Hall, B.A. (1978) Short-term validation of a selection procedure for student nurses for the general diploma in nursing. CSIR Special Report, PERS 318. Johannesburg: National Institute for Personnel Research.

Hall, B.A. (1980) Short-term validation of a selction procedure for student nurses for the general diploma in nursing. CSIR Special Report, PERS 318. Johannesburg: National Institude for Personnel Research.

Halstead, M.E. (1985) Development and validation fo three tests of arithmetic ability. CSIR Special Report, PERS 392. Johannesburg: National Institude for Personnel Research.

Hartigan, J.A., and Wigdor, A.K. (eds) (1989) Fairness in employment testing: validity generalisation, minority issues, and the General Aptitute Test Battery. Washington: National Academy Press.

Helmstadter, G.C. (1966) Principles of psychological measurements. London: Methuen.

Holburn, P. (1989) Apprentice selection: an HSRC/NTB survey of policies and methods used in the RSA with an emphasis on psychometric testing. CSIR Special Report, C/PERS 296. Johannesburg: National Institute for Personnel Research.

Holburn, P. (1990) Selection decisions: The quest for fairness. CSIR Special Report, PERS 424. Johannesburg: National Institute for Personnel Research.

Holburn, P. (Unpublished manuscript) Item bias research on four apprentice selection tests.

Hoorweg, J. (1976) Africa (South of the Sahara). In V S Sexton and H Misiak (Eds). Psychology around the world. Monterey, California: Brooks/Cole Publishing Co.

Hudson, W. (1962) National Institute for Personnel Research, 1946 - 1961. Psychologia Africana. 9, 13- 21.

Hunter, J.E. Schmidt, F.L., and Jackson, G.B. (1982) Advanced meta-analysis: Quantative methods for cumulating research findings across studies. San Francisco: Sage.

Huysamen, G.K. (1980) Psychological test theory. Pretoria: Academica.

Huysamen, G.K. (1981) Introductory statistics and research design for the behavioural sciences. Vol 1 Pretoria: Academica.

Huysamen, G.K. (1988) Psychological measurement: An introduction with South African examples. Cape Town: Academica.

Irvine, S.H. (1969) Factor analysis of African abilities and attainments: constructs across cultures. Psychological Bulletin. 71 (1), 20 - 32.

James, L.R., Demaree, R.G., and Mulaik, S.A. (1986) A note on validity generalisation procedures. Journal of Applied Psychology. 71: 440 -450.

Jenson, A.R. (1980) Bias in mental testing. London: Methuem.

Kail, R., and Pellegrino, J.W. (1985) Human intelligence: Perspectives and prospects. New York: W.H. Freeman and Company.

Kriegler, S.M. (1988) Opleiding van Opvoedkundige Sielkundiges vir die bevordering van Geestesgesondheid in Suid-Afrika. South African Journal of Psychology. 18, 84 -90.

Kuder, G.F., and Richardson, M.W. (1937) The theory of estimation of test reliability. Psychometrica . 2, 151 - 160.

Latti, V.I. (1978) Investigation into the validity of selection tests for clerks, typists and tellers in a building society. CSIR Special Report, C/PERS 267. Johannesburg: National Institude of Personnel Research.

Lautenschlager, G.J., and Mendoza, J.L. (1986) A step-down hierarchical multiple regression analysis for examining hypotheses about test bias in prediction. Applied Psychological Measurement. 10, 133 - 139.

Lawshe, C.H. (1975) A quantitative approach to content validity. Personnel Psychology. 28, 563 - 375.

Lewis, W.M. (1980) Validation of a test battery for the selection of chargehands. CSIR Special Report, C/PERS 296. Johannesburg: National Institute for Personnel Research.

Muchinsky, P.M. (1986) Psychology applied to work: An introduction to industrial and organisational psychology. Chicago: Dorsey Press.

Mauer, K.F., and Retief, A.I. (Eds). (1987) Psychology in context: cross-cultural trends in South Africa. Pretoria: Human Sciences Research Council.

National Institute for Personnel Research (1975) Test administrators manual for the High Level Battery (Rev. Ed.). Johannesburg: Council for Scientific and Industrial Research.

Neave, H.R. (1978) Statistics tables for mathematicians, engineers, economists and the behavioural and management sciences. London: George Allen and Unwin.

Nunnally, J.C. (1976) Psychometric theory. New York: McGraw-Hill.

Otis, A.S. (1073) Otis Quick Scoring Mental Ability Tests: Manual of directions for Beta Test forms A and B. New York: World Book Company.

Owen, K. (1986) Toets-en-itemsydigheid: Toepassing van Senior Aanlegtoets, Meganiese Insigtoets en Skolatiese bekwaamheidstoetsbattery op blanke,swart,kleuring-en Indier technikonstudente. IPER Research Report. Pretoria: Human Science Research Council.

Owen, K. (1989) Aptitude tests. In K Owen, and J J Taljaard (Ed.). Handbook for the use of psychological and scholastic tests of IPER and the NIPR. Pretoria: Human Sciences Research Council.

Pearlman, K., Schmidt, F.L., and Hunter, J.E. (1980) Validity generalisation results for tests used to predict job proficiency and training success in clerical occupations. Journal of Applied Psychology. 65, 373 - 406.

Peterson, J. (1925) Early conceptions and tests of intelligence. New York: World Book Company.

Pintner, R. (1924) Intelligence testing: Methods and results. London: University of London Press.

Poortinga, Y.H. (1971) Cross-cultural comparison of maximum performance tests: some methodological aspects and some experiments with auditory and visual stimuli. Psychologia Africana. Monograph Supplement No.6.

Poortinga,Y.H. (1975) Limitations on intercultural comparison of psychological data. Nederlands Tijdschrift voor de Psychologie.30,23 - 39.

Pottas, C.D., Erwee, R., Boshoff, A.B., and Lessing, B.C. (1980) Manual for the Achievement Motivation Questionnaire. Unit for Entepreneurship, University of Pretoria.

Prinsloo, R.J. (1982) The control of psychological tests by the Test Commission of the Republic of South Africa (TCRSA). South African Journal of Psychology. 12, 106 - 110.

Retief, A. (1986) The need for theory development in psychology. Some case studies. South African Journal of Psychology. 16, 71 - 78.

Retief, A. (1988) Method and theory in cross-cultural psychological assessment. Research Report Series: 6. Pretoria: Human Science Research Council.

Schmidt, F.L., Hunter, J.E., and Urry, V.W. (1976) Statistical power in criterion - related validity studies. Journal of Applied Psychology. 61, 473 - 485.

Schmidt, F.L., and Hunter, J.E. (1977) Development of a general solution to the problem of validity generalisation. Journal of Applied Psychology. 62, 529 - 540.

Schmidt, F.L. Hunter, J.E. (1978) Moderator research and the law of small numbers. Personnel Psychology. 31, 215 -232.

Schmidt, F.L., Hunter, J.E., McKenzie, R.S. and Muldrow, T.W. (1979) The impact of valid selection procedures on workforce productivity. Journal of Applied Psychology. 64, 609 - 626.

Schmidt, F.L., Hunter, J.E., Pearlman, K., and Shane, G.S. (1979) Further tests of the Schmidt-Hunter Bayesian validity generalisation procedure. Personnel Psychology. 32, 257 -281.

Schmidt, F.L., Gast-Rosenberg, I., and Hunter, J.E. (1980) Validity generalisation results for computer programmers. Journal of Applied Psychology. 65, 643 -661.

Schmidt, F.L., and Hunter, J.E (1980) The future of criterion-related validity. Personnel Psychology. 33, 41 - 60.

Schmidt, F.L., Hunter, J.E., and Caplan, J.R. (1981) Validity generalisation results for two job groups in the petroleum industry. Journal of Applied Psychology. 66, 261 - 273.

Schmidt, F.L., Hunter, J.E., and Pearlman, K. (1982) Progress in Validity Generalisation: Comments on Callender and Osburn and future developments. Journal of Applied Psychology. 67: 835 - 845.

Selltiz, C., Jahoda, M., Deutsch, M., and Cook, S.W. (1959) Research methods in social relations (2nd ed) New York: Holt, Rinehart and Wilson.

Siegel, S. (1956) Nonparametric statistics for the behavioural sciences. London: McGraw-Hill.

Spence, B.A. (1982) A psychological investigation into the characteristics of higher primary school guidance teachers. CSIR Special Report, C/PERS 332. Johannesburg: National Institute for Personnel Research.

Taylor, H.C., and Russell, J.T. (1939) The relationship of validity coefficient to the practical effectiveness of tests in selection. Discussion and tables. Journal of Applied Psychology. 23, 565 - 5.

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

Taylor, J.M., and Radford E.J. (1986) Psychometric testing as an unfair labour practice. South African Journal of Psychology. 15: 19 - 86.

Taylor, J.M. (1987) Fair selection practice for equal employment organisations. IPM Journal. 8 - 11

Taylor, T.R., Werbeloff, M., and Ebertsohn, M.R. (1982) The programming and validation of the computerised version of the intermediate Mental Alertness Test. CSIR Special Report PERS 335 Johannesburg: National Institude for Personnel Research.

Taylor, T.R. (1987) Test bias: the roles and responsibilities of test user and test publisher. CSIR Special Report, PERS 424. Pretoria: Human Sciences Research Council.

Taylor, T.R (1990, June) Are you testing fairly? Paper presented at a breakfast seminar, Durban.

Thorndike, R.L., and Hagen, E. (1969) Measurement and evaluation in psychology and education (3rd Ed). New York: John Wiley.

Tucker, L.R. (1949) A note on the estimation of reliability by the Kuder-Richardson Formula 20. Psychometrika. 14: 117 - 119.

Tyler, L.E. (1971) Test and Measurements (2nd ed). New Jersey: Prentice-Hall.

Van den Berg, A.R. (1985). Using the Junior South African Individual Scale (JSAIS) for testees from South African population groups which were not included in the norm population. Institute for Psychometric and Edumetric Research, Catalogue No. 2385, Pretoria: Human Science Research Council.

Van den Berg, A.R. (1989) How to estimate and evaluate test reliability. Pretoria: Human Sciences Research Council.

Van de Vijver, F.J., and Poortinga, Y.H. (1982) Corss-cultural generalisation and universality. Journal of Cross-Cultural Psychology. 13: 387 - 498.

Van Staden, F.J. (1987) A decade of Environmental Psychology in South Africa. South African Journal of Psychology. 17, 72 - 75.

Van Staden, F., and Visser, D. (1990) Analysis of themes and statistical techniques: A review of the past decade of the South African Journal of Psychology. South African Journal of Psychology. 20,(1).

Van Zyl, T., and Myburgh. C. (1991) HSRC cataloque part 1: Psychological and proficiency tests: Individual and group intellegence tests, personality tests and aptitude and proficiency tests. Pretoria: Human Sciences Research Council.

Verster, J.M. (1987a) Cross-cultural cognitive research: Some methodological problems as prospects. In K F Mauer and A I Retief (Eds). Psychology in context: Cross-Cultural research trends of South Africa. Pretoria: Human Sciences Research Council.

Verster, J.M. (1987b) Speed of cognitive processing: Cross-culture findings on structure and relation to intelligence, tempo, temperament and brain function. Paper presented to NATO Advanced Study Institude on Human Assessment, Athens, Greece. Johannesburg: National Institude for Personnel Research, Human Science Research Council.

Visser, B.L. (1977) Vocational counselling at the NIPR. CSIR Special Report, PERS 257. Johannesburg: National Institute for Personnel Research.

Visser, B.L. (1978) Development of a testing programme for Black university students for purposes of selection and careers counselling. CSIR Special Report, PERS 281. Johannesburg: National Institude for Personnel Research.

Visser, B.L. (1980) Electroencephalographic assessment in vocational counselling. CSIR Special Report, PERS 296. Johannesburg: National Institude for Personnel Research.

Werbeloff, M., and Taylor, T.R. (1982) Development and validation of the High Level Figure Classification Test. CSIR Special Report, PERS 338. Johannesburg: National Institude for Personnel Research.

Whipple, E. (1910) Manual of mental and physical tests. Baltimore: Warwick and York.

Wilcocks, A.M. (1973) Intermediate Battery test administrator's manual. Johannesburg: National Institute for Personnel Research.

Wober, M. (1975) Psychology in Africa. London: Internal African Institute.

Yates, A. (1959) Review of the Otis Quick-Scoring Mental Ability Tests. In O.K. Buros (ed.) Fifth Mental Measurement Yearbook. New Jersey: Gryphon.